Overview of b tagging at CMS

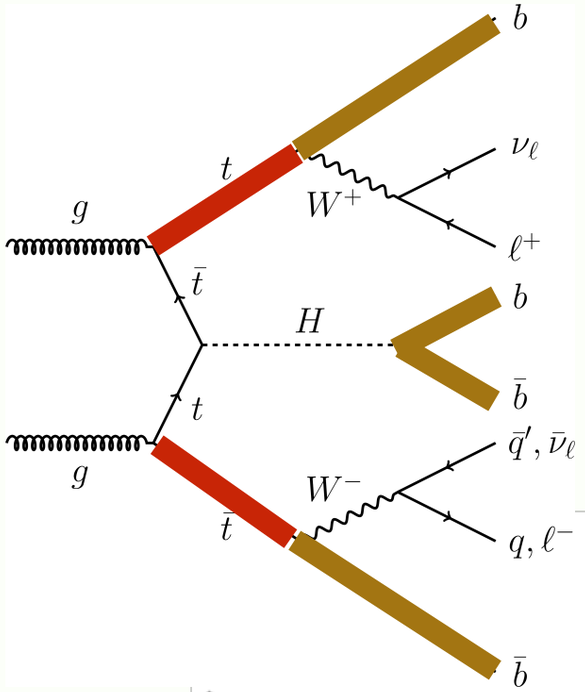
Kirill Skovpen (IPHC Strasbourg)

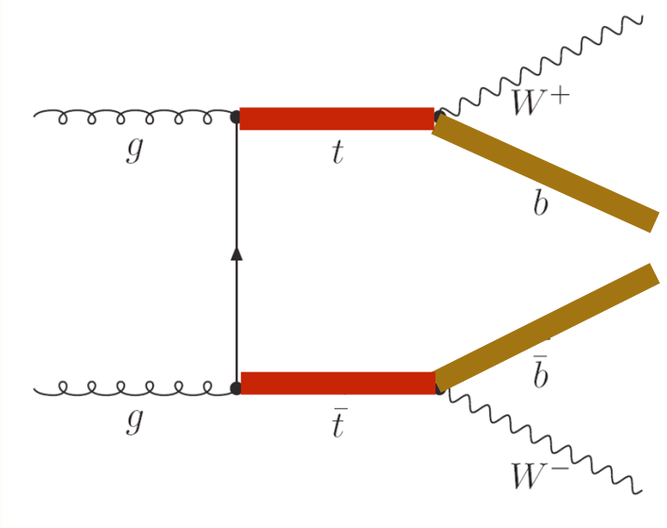
Top LHC France
Clermont-Ferrand

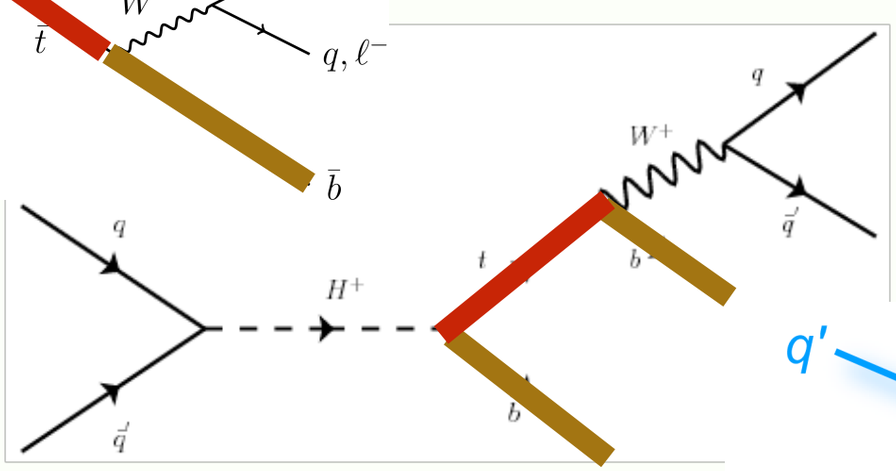
May 19, 2016

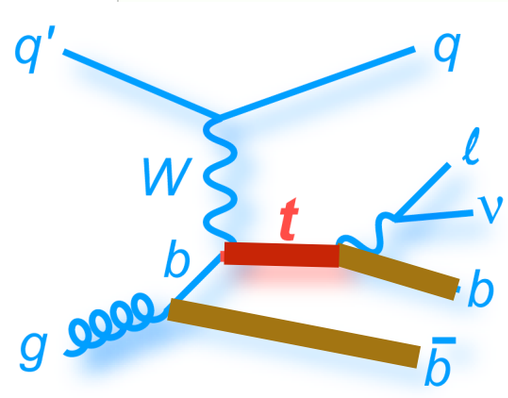
Outline and motivation

- Algorithms, commissioning and trigger
- Performance measurements
- Upgrade studies

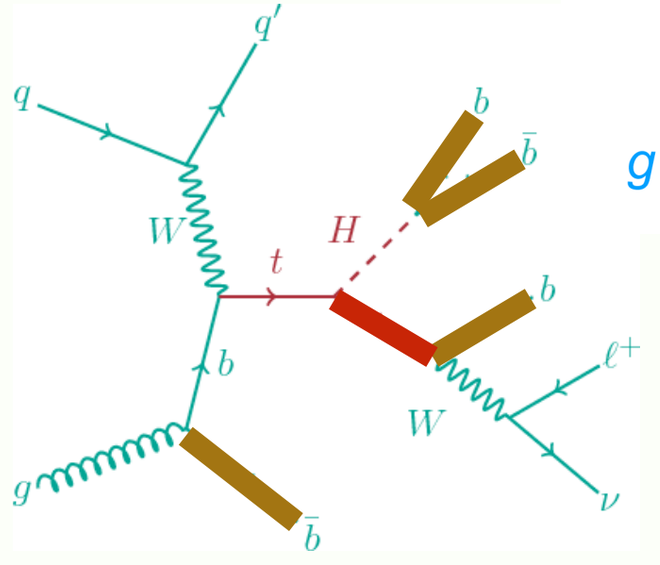








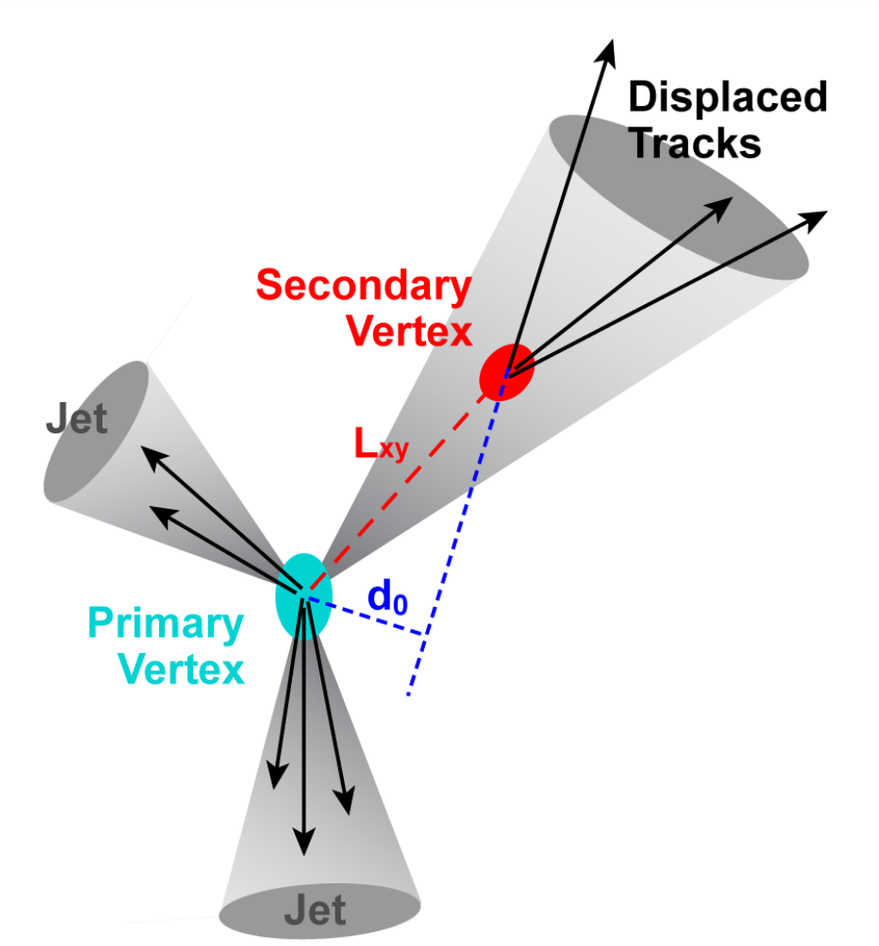
- **b quarks** are always present in **top quark** decays with Br(t→Wb) ≈100%
- top quark production could also be accompanied by b-quarks
- b jet identification (***b tagging***) is **crucial for Top SM and BSM physics** !



Introduction to b tagging

- **b jets** = jets that arise from the process of hadronization of b quarks
- Many physics analyses (Top, Higgs, Exotics) rely on efficient identification of b jets

- Use **B-hadron properties** to identify b jets:
  - Relatively large mass [5-6 GeV]
  - Long lifetime [$c\tau \approx 450$ μm]
    E = 70 GeV gives $\beta\gamma c\tau \approx 5$ mm
  - Daughter particle multiplicity
    ≈ five charged tracks per decay
  - Possible presence of semileptonic decays
    $b \rightarrow \mu\nu X$ [Br ≈ 11%], $b \rightarrow c \rightarrow \mu\nu X$ [Br ≈ 10%]
  - Tertiary vertex
    (B-meson decay to a charmed hadron), $c\tau \approx$ 120-310 μm



Algorithms, commissioning and trigger

b tagging algorithms

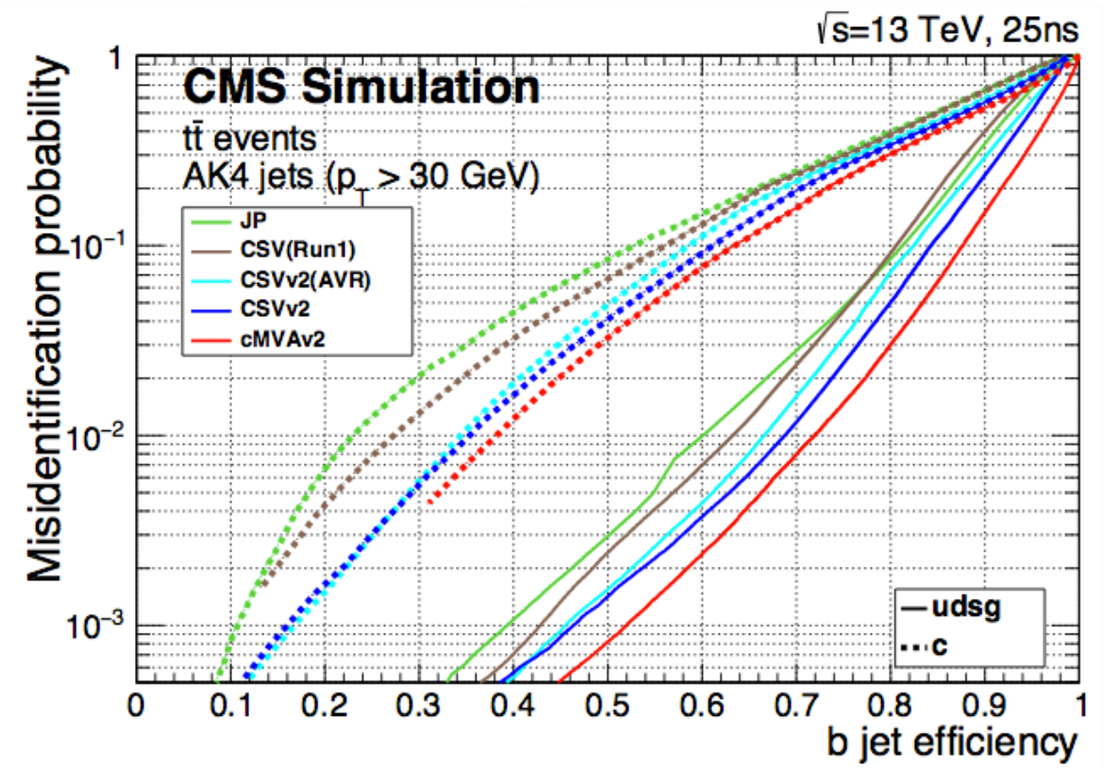
Algorithm	ATLAS	CMS
Impact parameter based	IP2D, IP3D, TrackCounting, JetProb	TCHP, TCHE, JP, JPB
Secondary vertex based	SV0, SV1, SV	SSVHP, SSVHE
Decay chain multi-vertex	JetFitter	
Soft lepton	SMT, p_TRel	Soft Lepton Taggers
Multivariate	JetFitterCombNN, MV1c, **MV2c00**, **MV2c20**	CSV, **CSVv2**, **cMVAv2**

Operating points either based on b-tagging or mis-tagging efficiencies:

b-tag: 60%, 70%, 77%, 85% (ATLAS)
mis-tag: 0.1%, 1%, 10% (CMS)

Flagship b taggers for Run 2

CMS b taggers



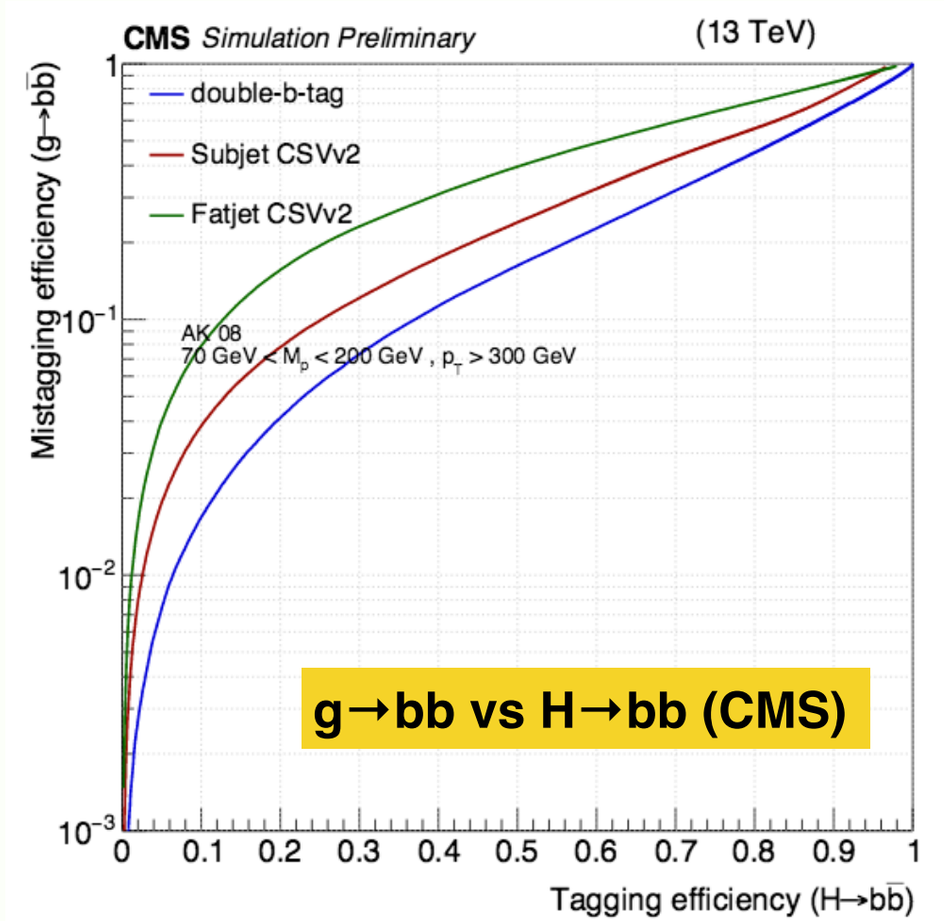
- **cMVAv2** is a new best b tagger - la grande **MVA-based combination of all taggers**
- SoftLepton information is used in cMVAv2 - unbiased calibration is only possible in ttbar events
- The new version of **CSVv2** for Run 2 is still widely used at CMS
- CSVv2 uses the **Inclusive Vertex Finder** algorithm instead of Adaptive Vertex Reconstruction (AVR) used in Run I - ***improved identification efficiency*** in boosted topologies and ttbar events

BTV results for 2015 13 TeV data: CMS-PAS-BTV-15-001

b tagging in boosted topologies

- **More boosted objects** with the increase of total energy
- b quarks could be present in decays of boosted particles
- Decay products ***clustered in a single fat (large-R) jet***
- Use ***jet substructure techniques*** to reconstruct sub jets and apply b tagging
  - Resolve b jets from top quark and Higgs decays, gluon splitting
  - Different signatures define different methods:
    - **top tagging**
    - **Double b tagging**

- **Dedicated tagger** to ***tag X→bb*** events
  - Improve discrimination against g→bb
  - Trained on G*→hh→4b vs QCD

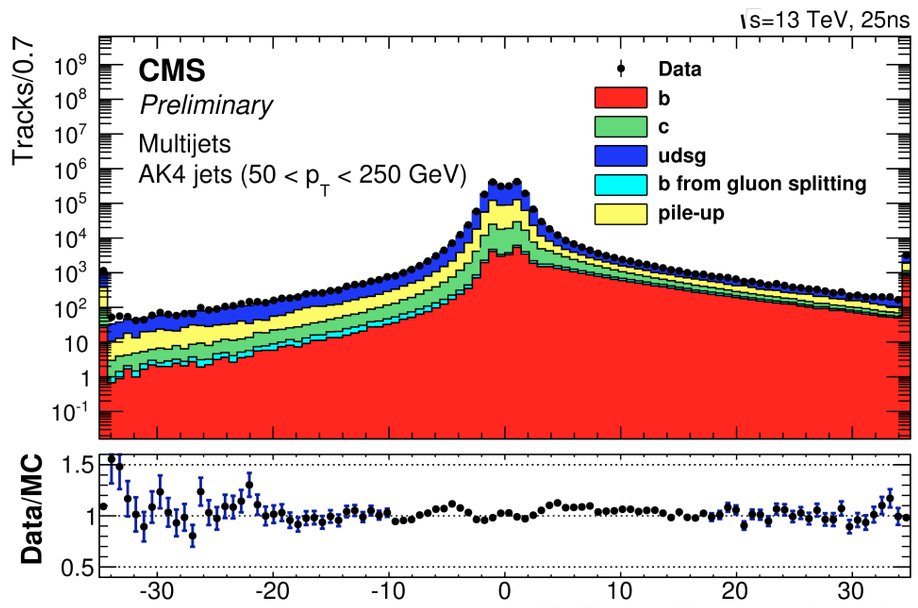


CMS DP-2015/038

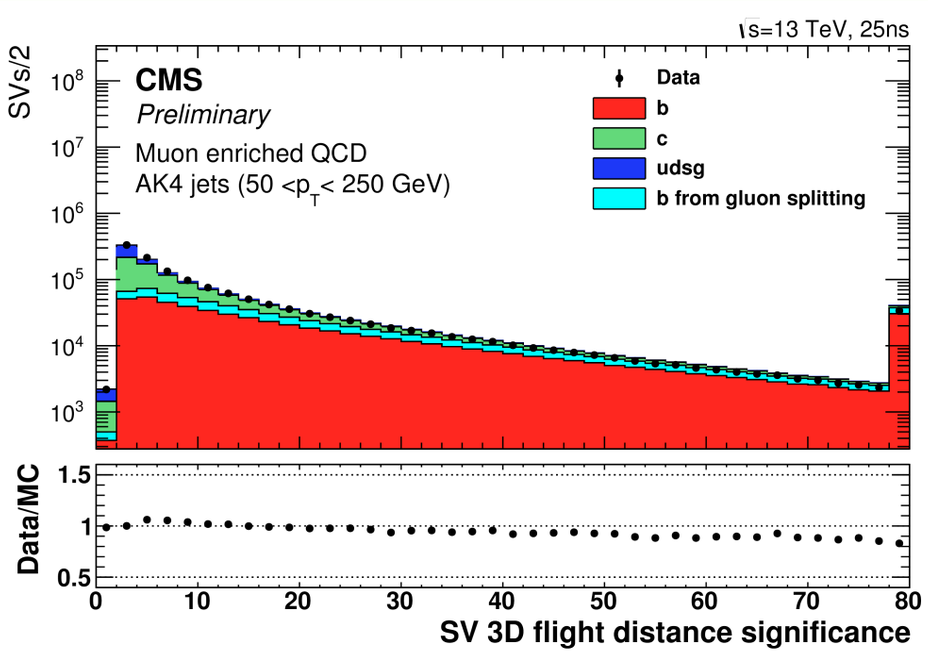
Commissioning for AK4 b jets

Validate the data/MC agreement for b tagging variables

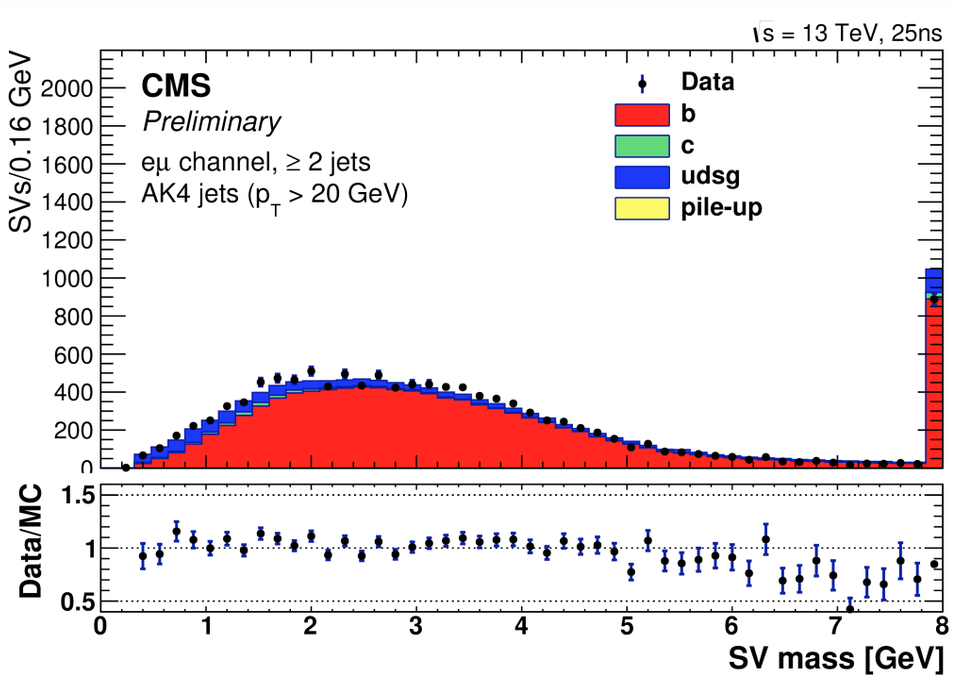
QCD inclusive: all jets



QCD muon-enriched: jets with muons



ttbar dilepton: all jets

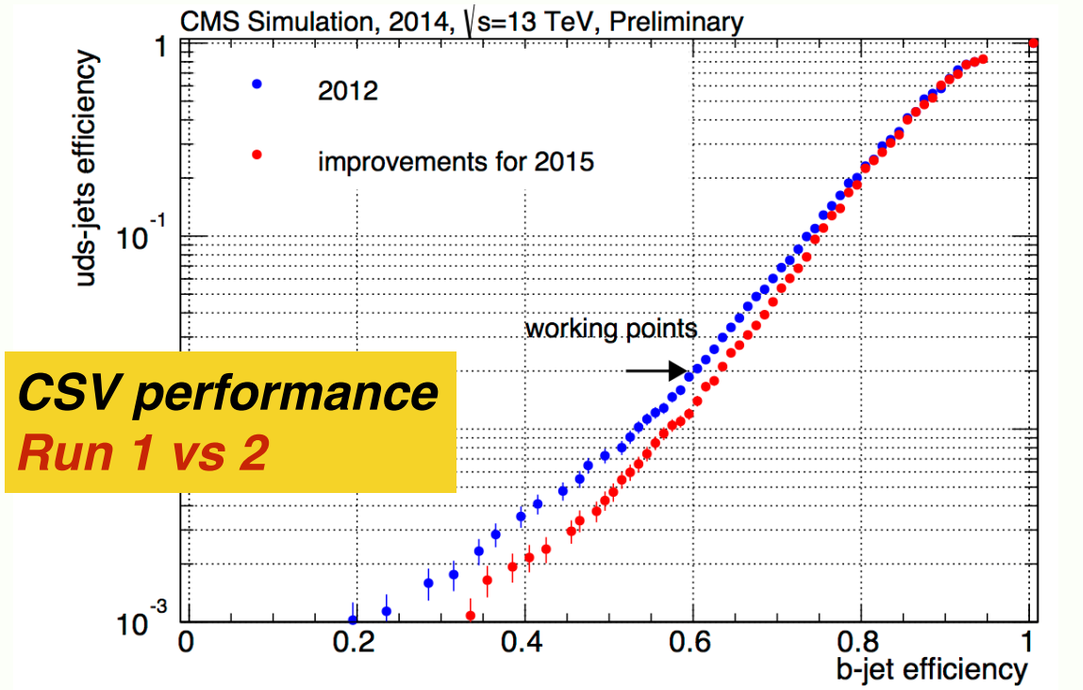


b jet trigger in Run 2

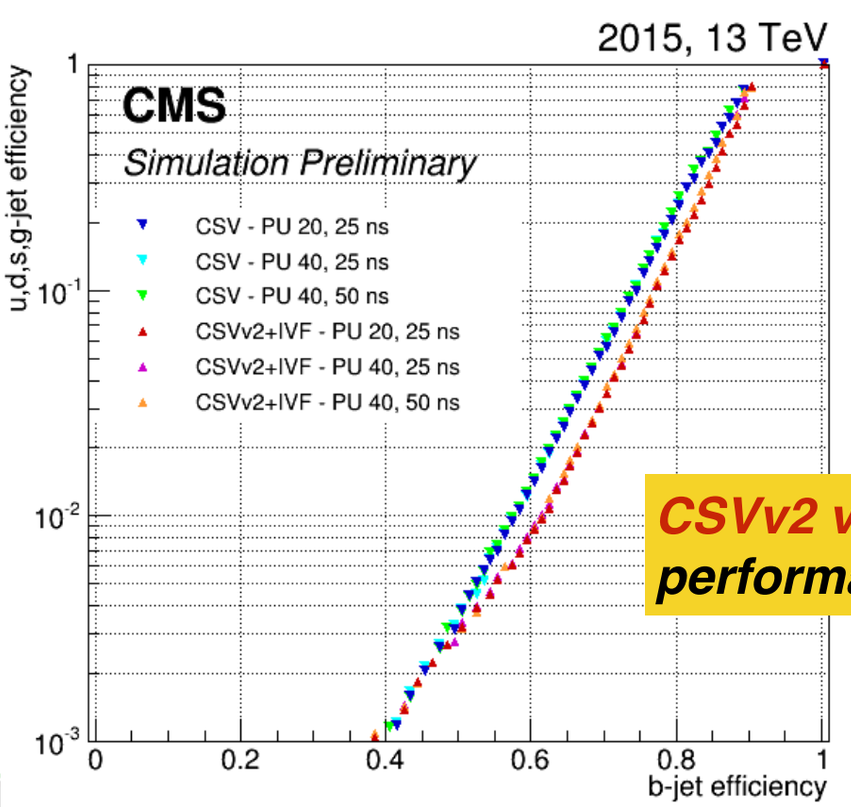
- ***L1*** is a ***hardware***-based trigger [output rate ≈**100 kHz**]
- ***HLT*** is a ***software***-implemented trigger executed on a multi-processor farm [output rate ≈**1 kHz**]
  - Many analyses that rely on b-jet identification (VBF H(bb), Z(νν)H(bb), HH(4b), etc.) ***suffer from very high trigger rates*** due to QCD multi-jet production
    - ***Implement b tagging selection at trigger level*** to significantly reduce the rate
    - b jet triggers already extensively used during Run I with several important upgrades for Run 2

HLT-based b-jet trigger

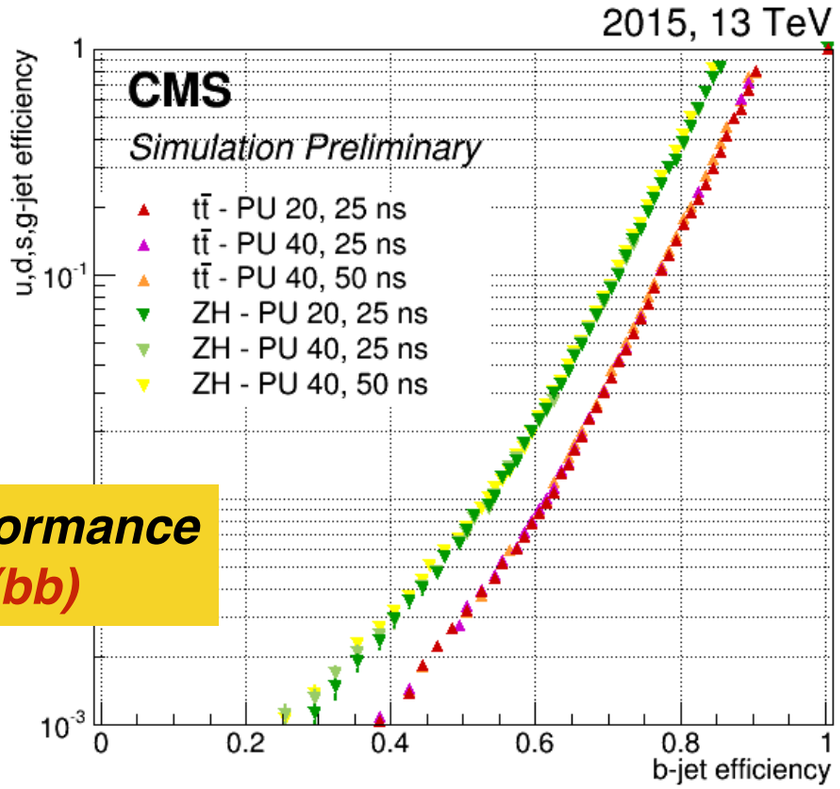
b jet trigger performance at CMS



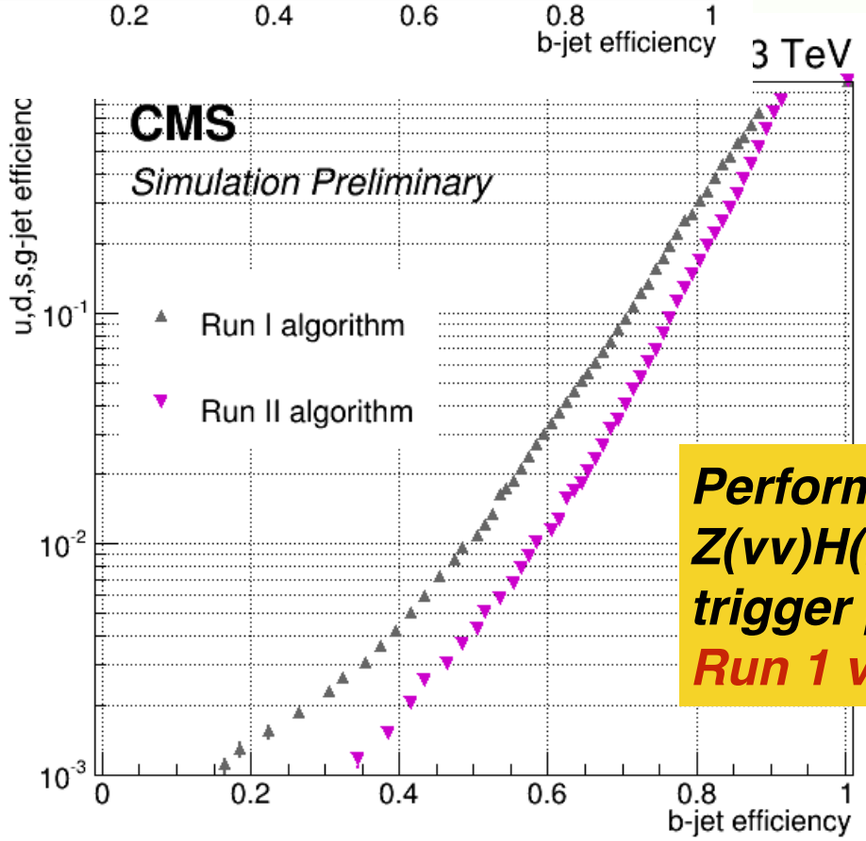
CSV performance
Run 1 vs 2



CSVv2 vs CSV performance



CSVv2 performance
ttbar vs ZH(bb)



Performance of Z(vv)H(bb) trigger path
Run 1 vs 2

Performance measurements

Calibration of b tagging performance

Calibration of b tagging efficiencies in data and MC is usually done in:

- **QCD multijet** events with b jets containing muons
- **ttbar** events with inclusive jets

Sample	ATLAS	CMS
QCD	p_TRel, System8	p_TRel, System8, Lifetime Tag (LT)
ttbar	Tag counting, Kinematic selection, Kinematic fit, Combinatorial likelihood	Flavour tag consistency, Tag counting, Tag & probe, bSample, Flavour tag matching, LT, KIN

Eventually, a **combination** of measured data/MC b tag SFs from different methods is performed

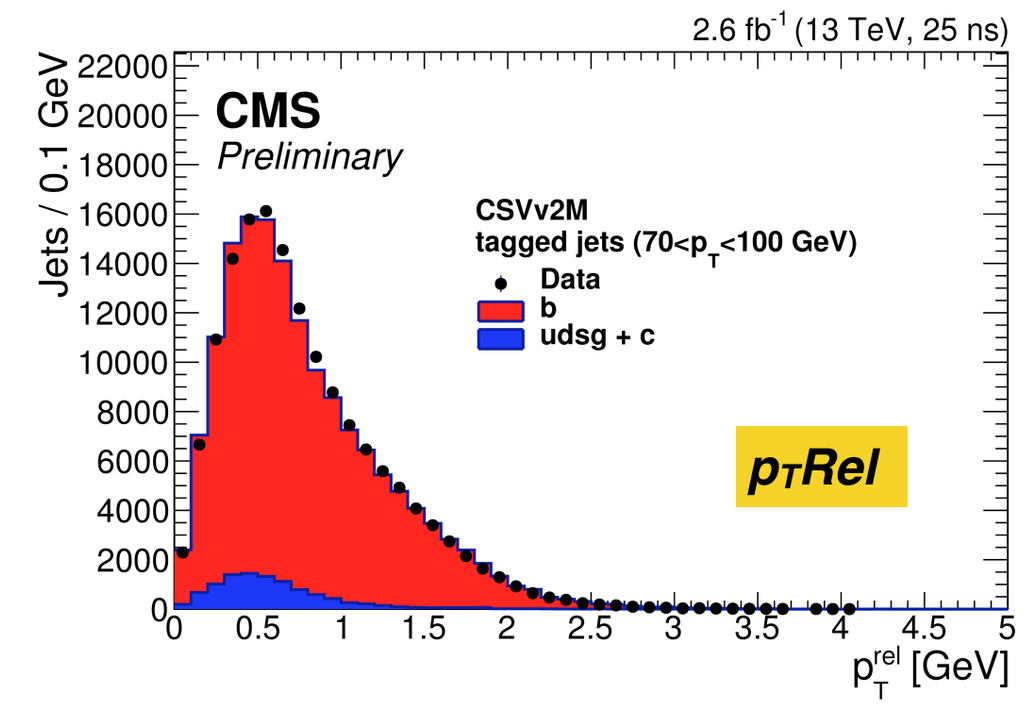
b tag calibration in QCD events

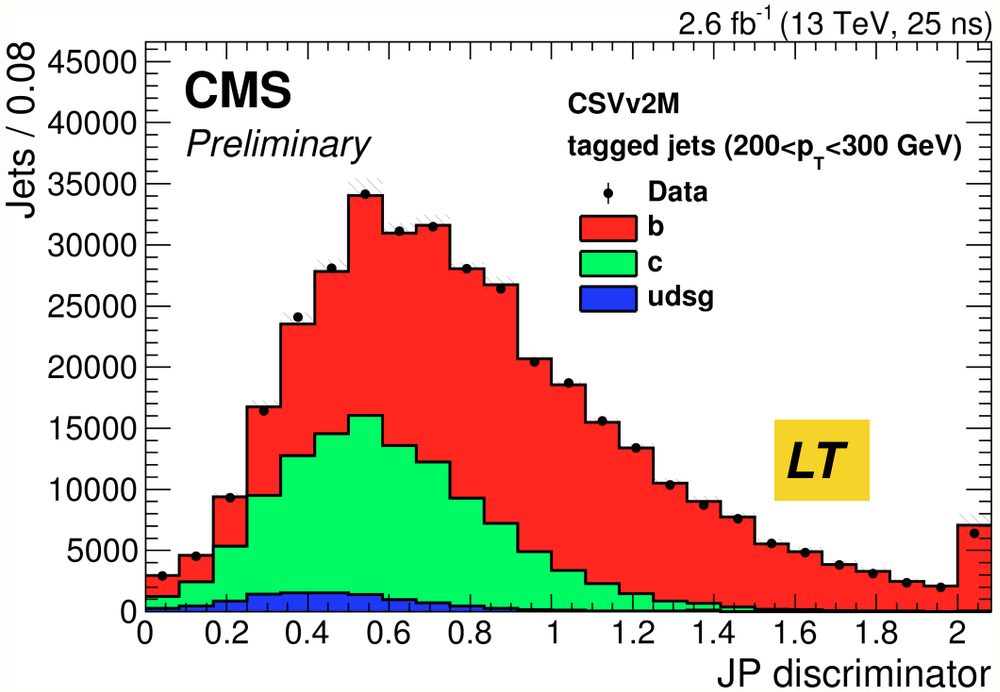
- **Template fit methods** based on ***p_TRel*** and Jet Probability (JP) discriminant (***LT***)
  - **p_TRel** - momentum of muon transverse to muon+jet axis
  - **JP** - compatibility of a set of tracks from a jet to originate from a primary vertex:

$$P_{tr}(S) = sign(S) \int_{|S|}^{\infty} R(x)dx$$

Resolution function built with negative IP tracks

- Fits are done before (or for jets failing b tagging) and after b tagging requirement to measure the efficiency

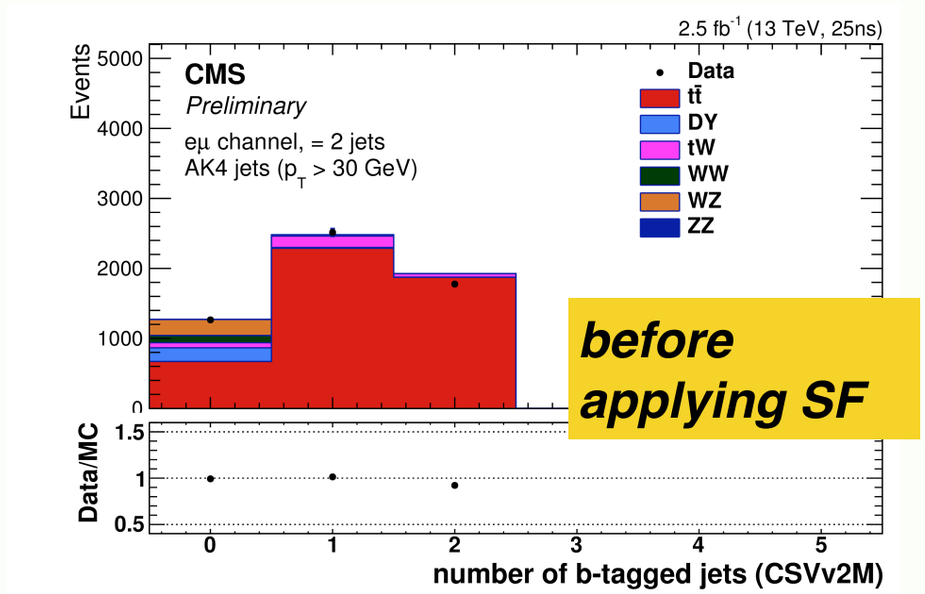


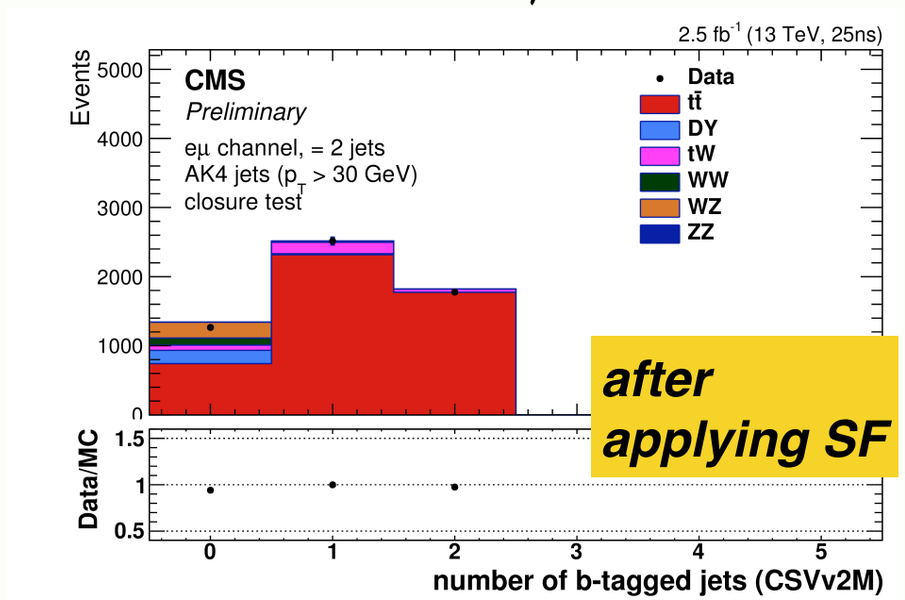


b tag calibration in ttbar events

- ***Tag counting method***
  - Use dilepton ttbar eμ events → high b jet purity
  - Count fraction of events with $N_{btag} = 2$ in a sample with two jets
  - Based on fractions → event yield systematics cancel out, but sensitive to modeling uncertainties (fragmentation and normalization scales)
  - No fit performed, calculate b tagging efficiency as:

$$\varepsilon_b = \sqrt{\frac{F_{2tag} - F_{non2b}^{truth}}{f_{2b}}}$$

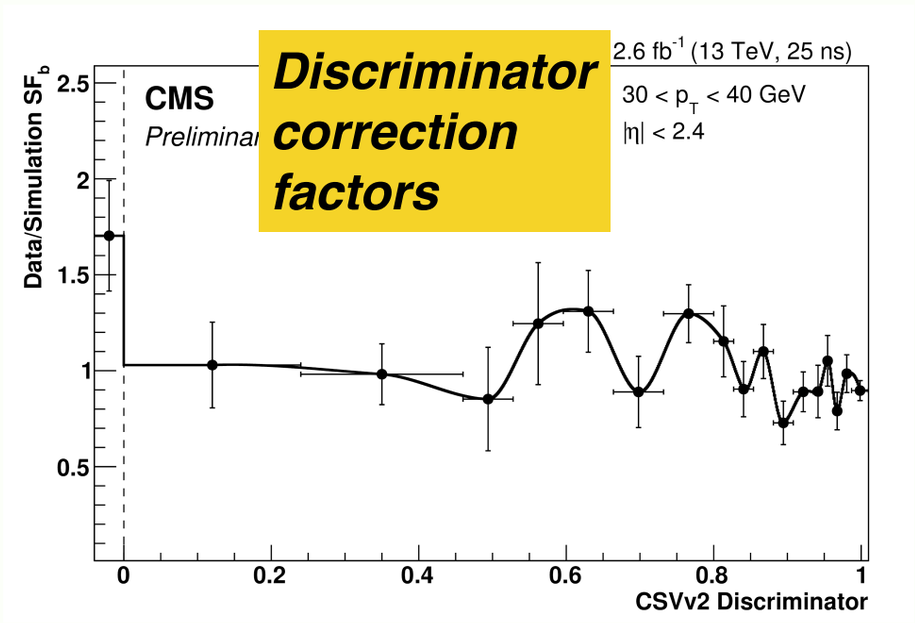




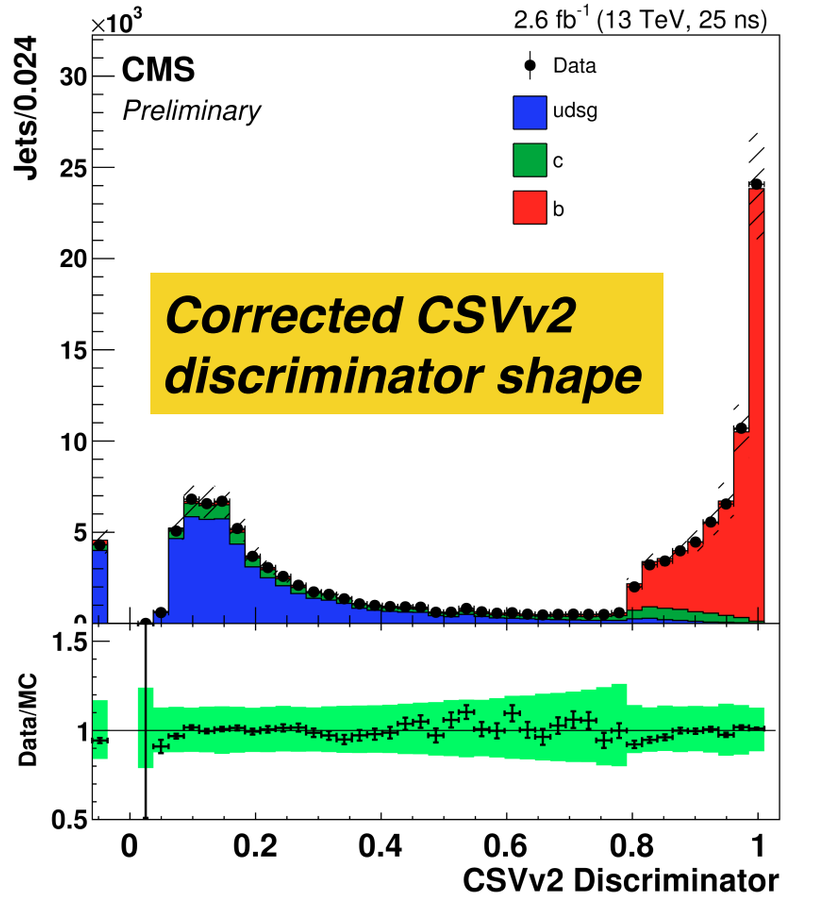
- ***Tag and probe method (see next slide)***

Shape correction for b tag discriminator

- ***b tag discriminator shape*** is often used in signal extraction techniques in many analyses
- Efficiency correction is needed over the whole range of discriminator values

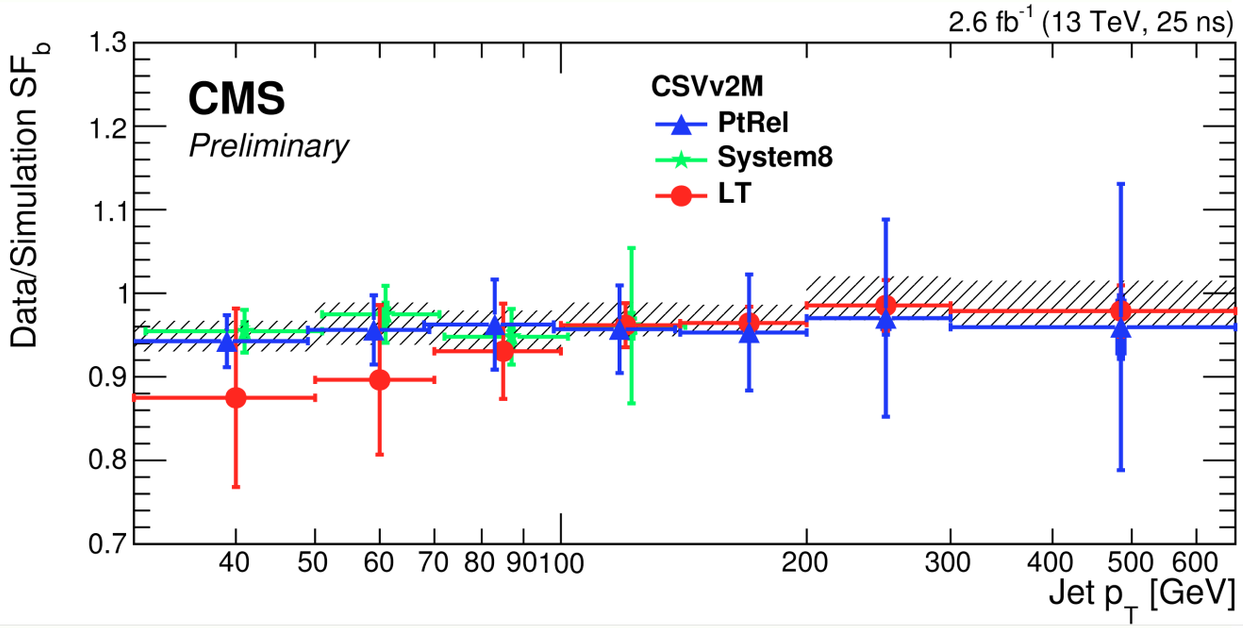


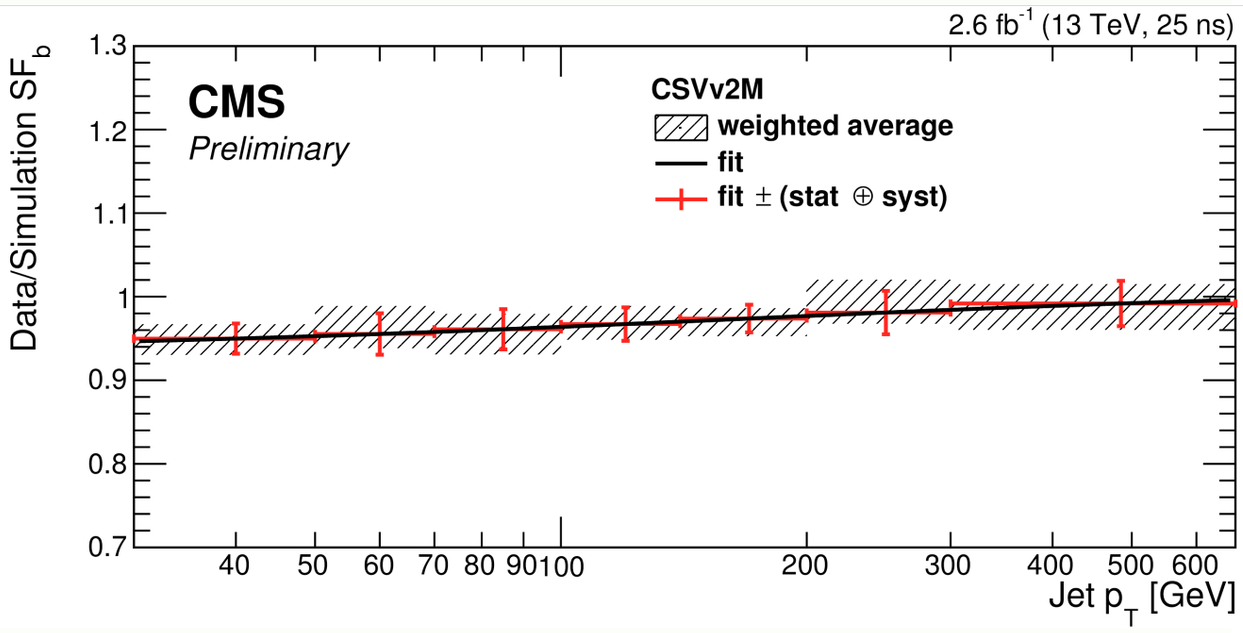
- **CMS** uses ***b tag re-weighting*** approach with correction factors measured with Tag&Probe method
  - **b-jet** in **ttbar dilepton** events
  - **l-jet** in **Z+jets**
- Also a shape correction via interpolation between the measured SFs is used



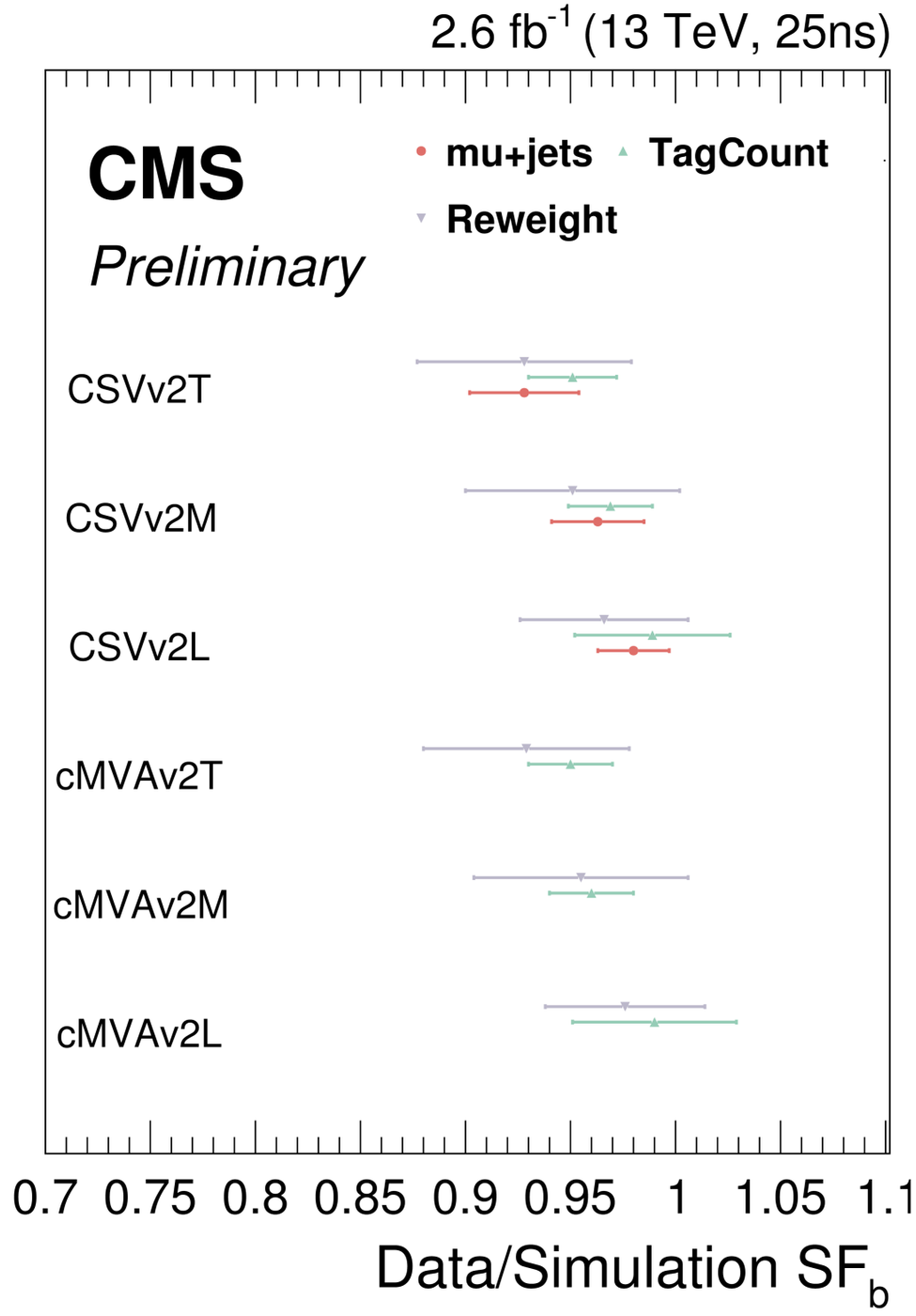
b jet scale factor results

QCD-based SF combination





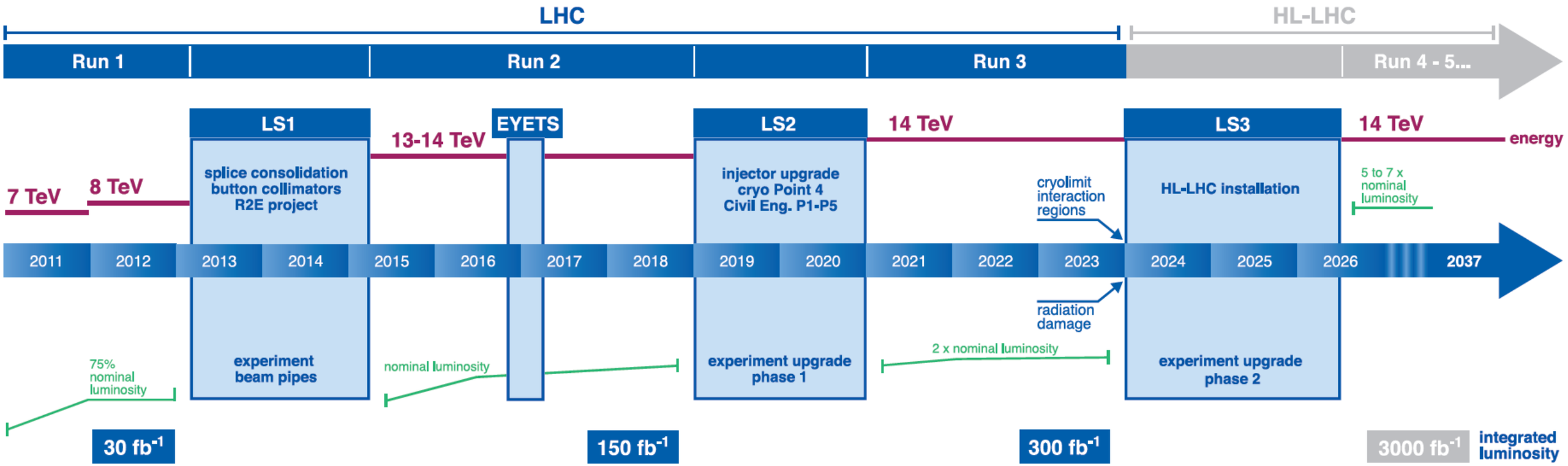
Comparison of QCD vs ttbar results



Upgrade studies

LHC roadmap

LHC / HL-LHC Plan



L = 0.75x10^{34}
50ns bunch
pileup ≈ 25

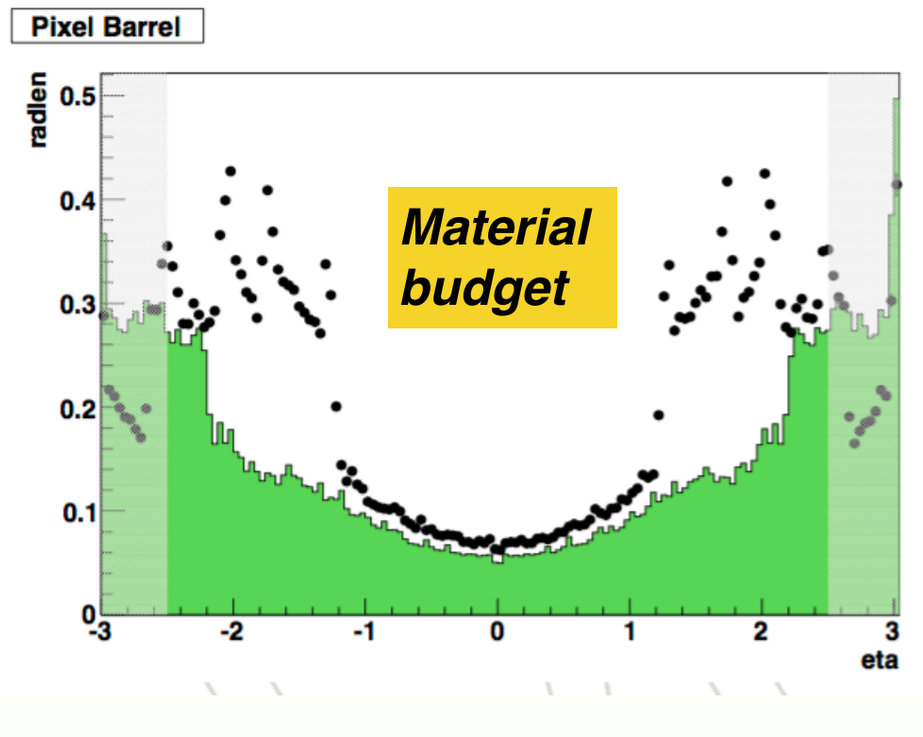
L = 1.5x10^{34}
25ns bunch
pileup ≈ 20-50

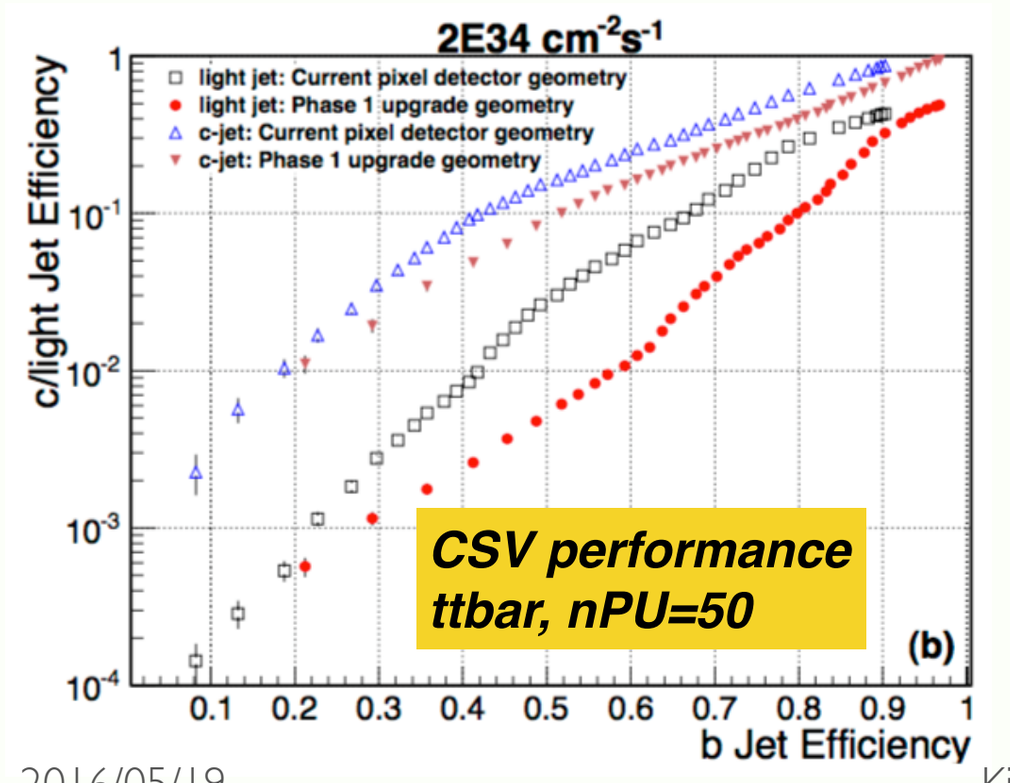
L = 1.7-2.2x10^{34}
25ns bunch
pileup ≈ 60

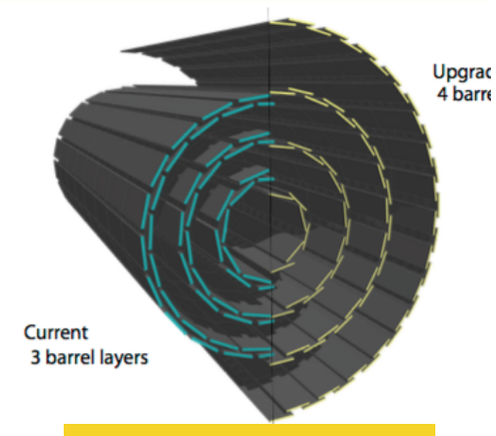
http://hilumilhc.web.cern.ch

CMS Phase I Pixel upgrade

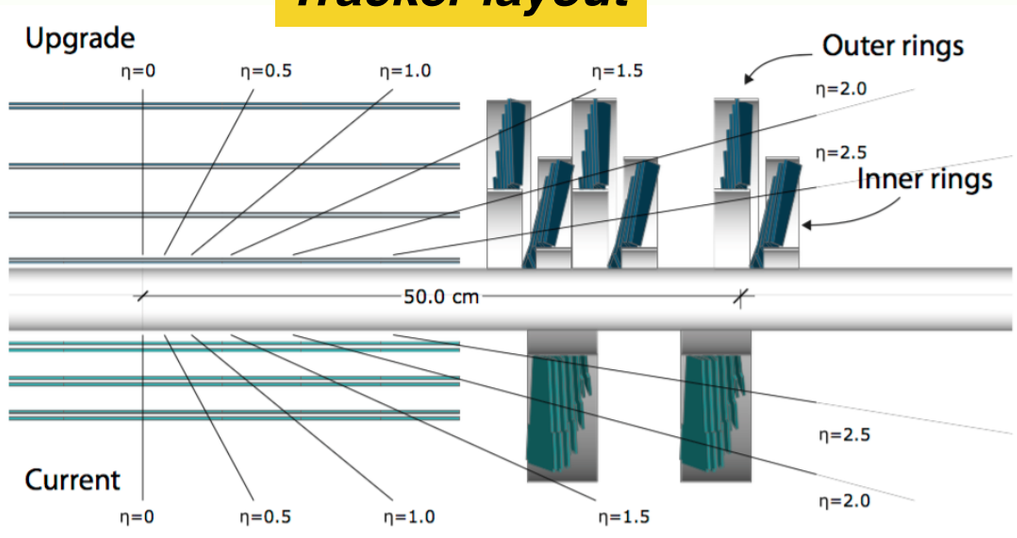
- The **present** pixel detector was designed for a luminosity of **10^{34}** cm^{-2} s^{-1} and pileup of 25 @25ns
- **After LS2**, luminosity to reach **$2x10^{34}$** and will result in the *large data losses in ROCs*
- The ***new beam pipe*** [R_{in} ≈ 22.5 mm] **installed in LS1**, ***full replacement of the pixel detector*** **by the end of 2016**
  - **Significant improvement in b tagging** due to extra layers, finer granularity, decrease in the amount of material





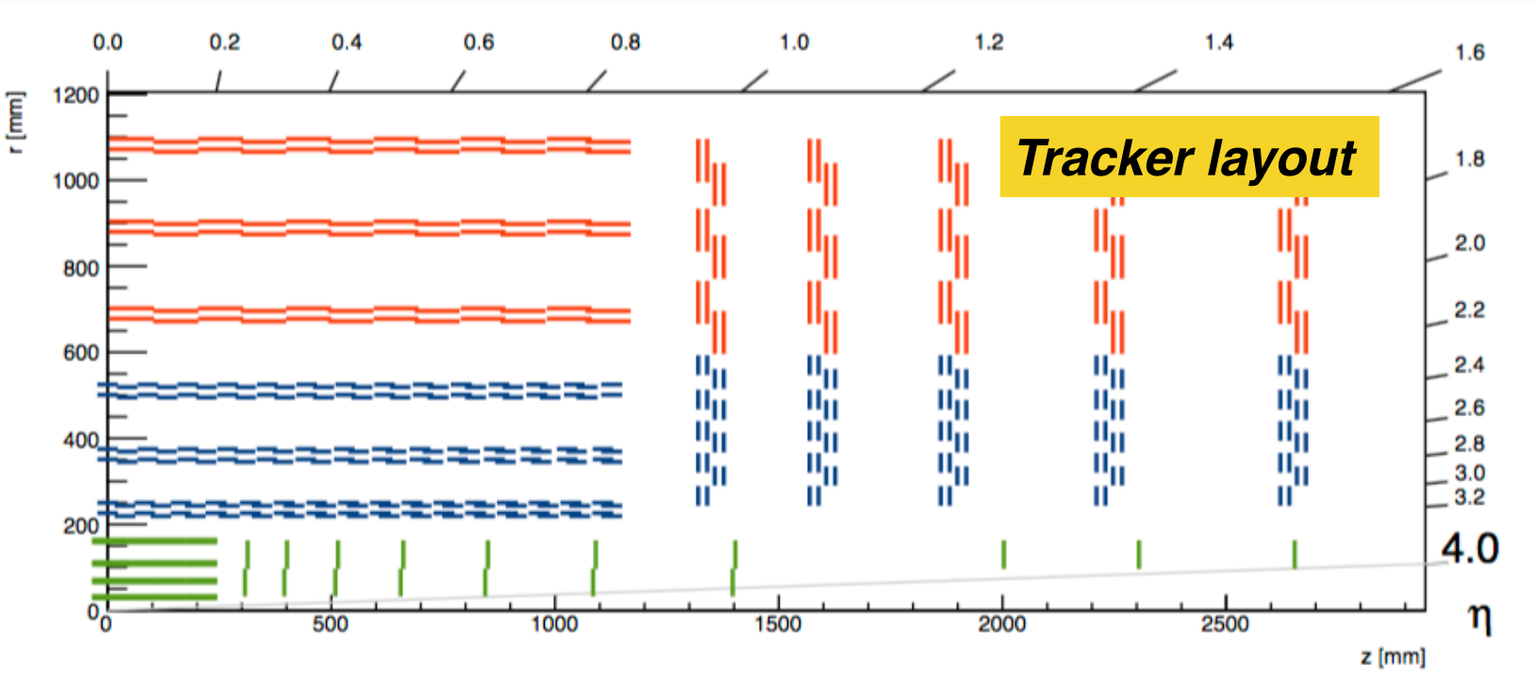


Tracker layout

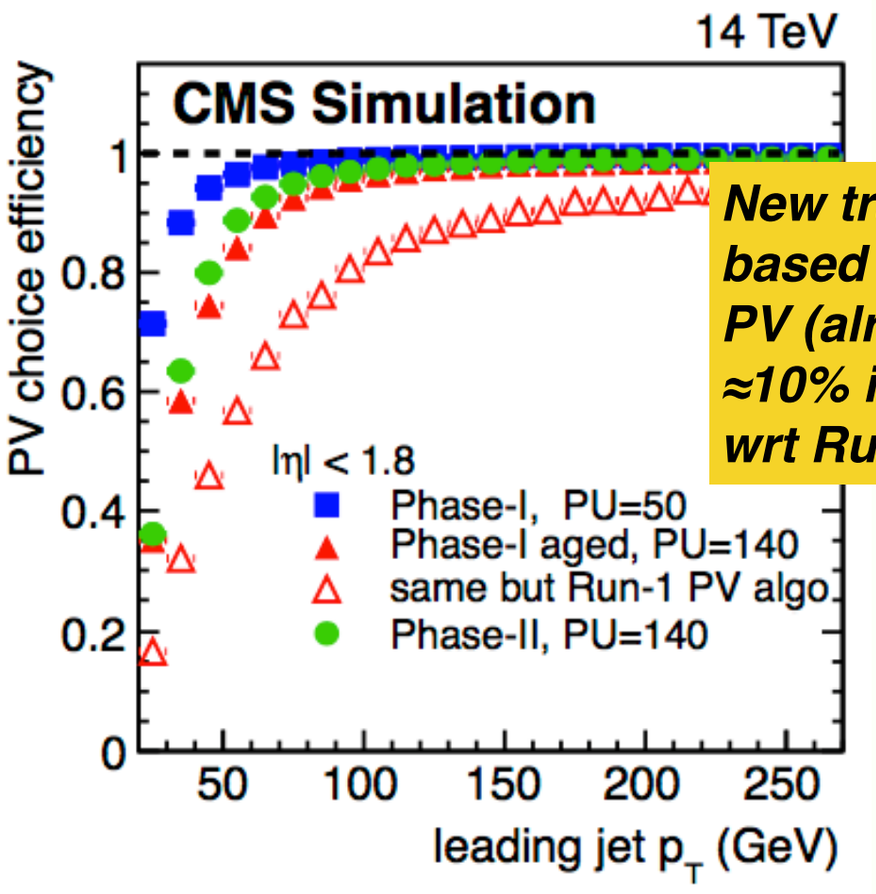


CMS-TDR-011
CERN-LHCC-2011-006

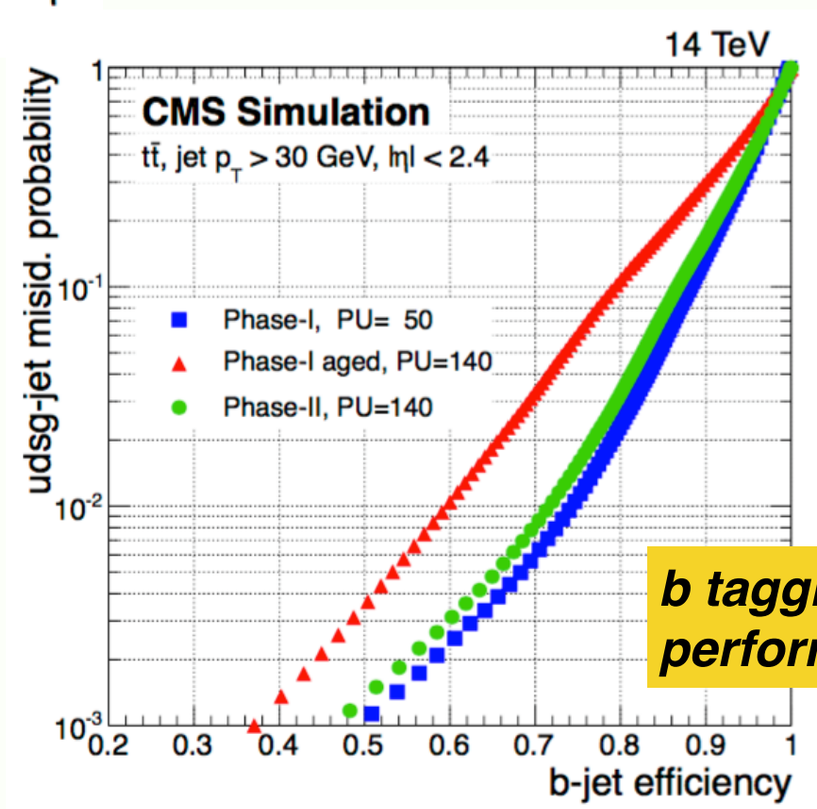
CMS Phase II Inner Tracker upgrade



- **Tracker to be completely replaced**
  - Finer granularity [≈4x]
  - L1 track trigger
  - Radiation hardness
  - b-tagging up to $|\eta| = 3$



New track-object-based algorithm for PV (already in Run 2), ≈10% improvement wrt Run 1



b tagging performance

CERN-LHCC-2015-010
CERN-LHCC-2015-019

Conclusion

- Last year brought the first 13 TeV data and **many new b tagging results at CMS**
- **Significant improvements** in b tagging efficiencies and reliable performance in **Run 2**
- New **c tagger** (BTV-16-001) and upgraded **double b tagger** (BTV-15-002) are coming soon
- **Very strong involvement of IPHC in BTV group activities at CMS**: Caroline (L2 BTV convener), Xavier (L3 b jet trigger convener), Anne-Catherine, Eric, Daniel (The Grand Master), Pierre (Mistag guru), Jérémy & Michaël (Top enthusiasts) and myself (L3 BTV performance convener)

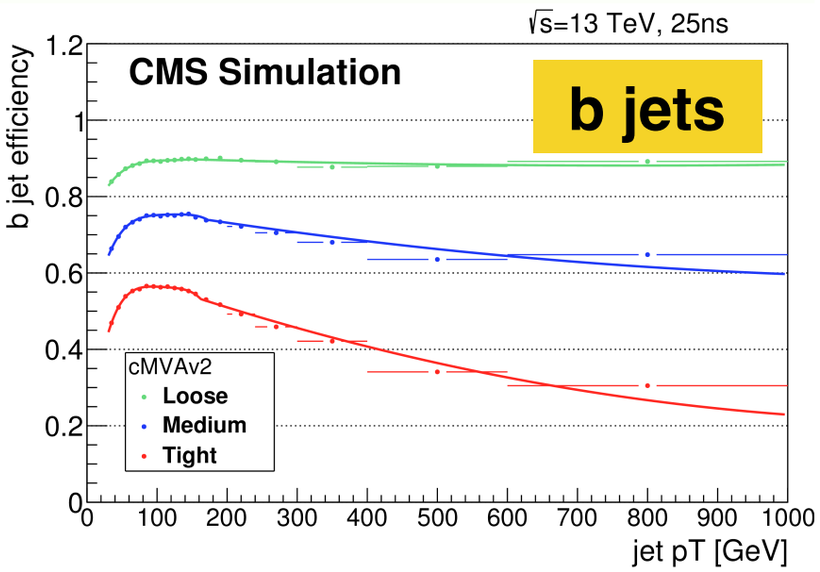
BACKUP

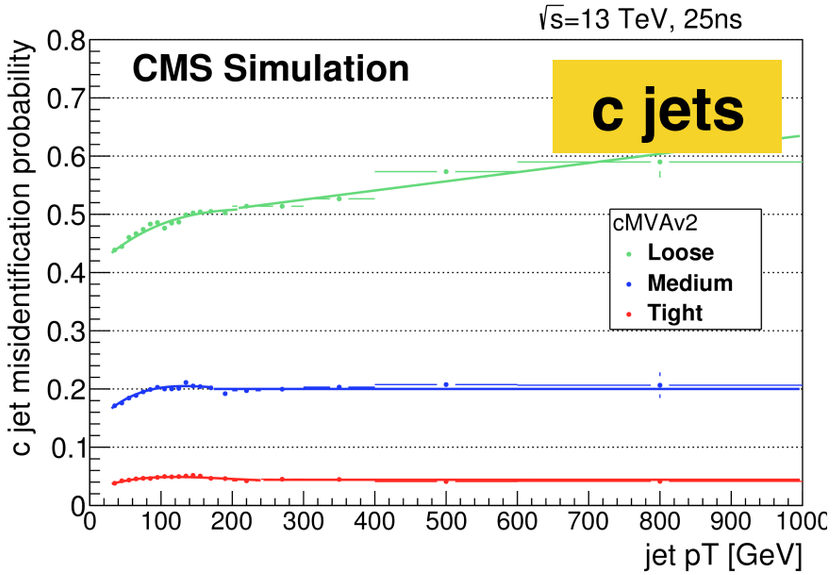
Challenges in Run 2

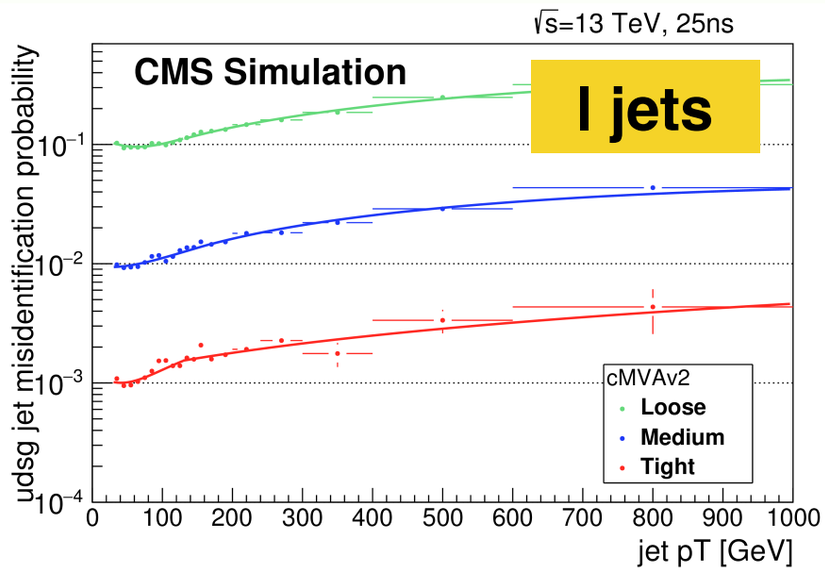
- **LHC beam collisions** setup during **Run 2** includes:
  - Higher center-of-mass energy of 13 TeV (was 8 TeV)
  - Higher instantaneous luminosity of 1.3×10^{34} cm^{-2} s^{-1} (was 7×10^{33})
  - Smaller bunch spacing of 25 ns (was 50 ns)

- **Challenges**:
  - Larger number of pileup interactions
    Not yet in 2015 but expected to be up to $\approx$ 40 by LS2
    Affects track and vertex reconstruction, dynamic inefficiency, increased occupancy
  - Increased trigger rates
  - Higher probability of boosted objects imposes requirements on tracking performance

b tagging efficiencies







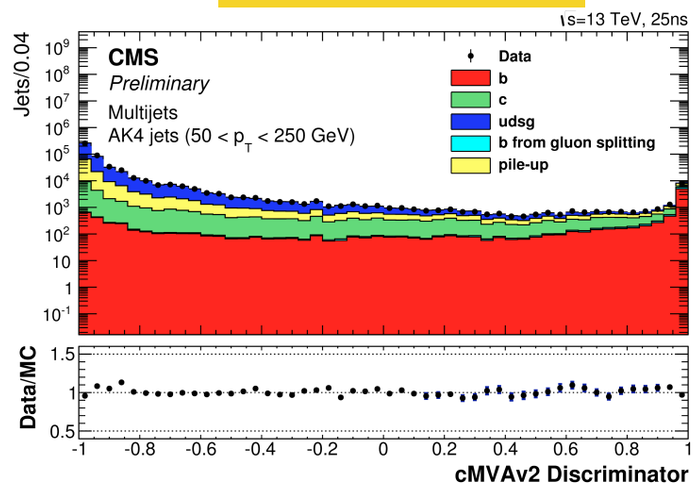
Jet flavour	operating point	jet p_T range	function
b	Loose	$30 \leq p_T < 150\,\text{GeV}$	$0.707 + 5.6 \cdot 10^{-3} \cdot p_T - 6.27 \cdot 10^{-5} \cdot p_T^2 + 3.10 \cdot 10^{-7} \cdot p_T^3 - 5.63 \cdot 10^{-10} \cdot p_T^4$
		$150 \leq p_T$	$0.906 - 6.39 \cdot 10^{-5} \cdot p_T + 4.11 \cdot 10^{-8} \cdot p_T^2$
	Medium	$30 \leq p_T < 175\,\text{GeV}$	$0.421 + 0.0107 \cdot p_T - 1.314 \cdot 10^{-4} \cdot p_T^2 + 7.268 \cdot 10^{-7} \cdot p_T^3 - 1.523 \cdot 10^{-9} \cdot p_T^4$
		$175 \leq p_T$	$0.79 - 3.17 \cdot 10^{-4} \cdot p_T + 1.24 \cdot 10^{-7} \cdot p_T^2$
	Tight	$30 \leq p_T < 160\,\text{GeV}$	$0.127 + 0.01578 \cdot p_T - 2.126 \cdot 10^{-4} \cdot p_T^2 + 1.273 \cdot 10^{-6} \cdot p_T^3 - 2.88 \cdot 10^{-9} \cdot p_T^4$
		$160 \leq p_T$	$0.634 - 6.74 \cdot 10^{-4} \cdot p_T + 2.69 \cdot 10^{-7} \cdot p_T^2$
c	Loose	$30 \leq p_T < 205\,\text{GeV}$	$0.40 + 1.23 \cdot 10^{-3} \cdot p_T - 4.60 \cdot 10^{-6} \cdot p_T^2 + 5.71 \cdot 10^{-9} \cdot p_T^3$
		$205 \leq p_T$	$0.478 + 1.573 \cdot 10^{-4} \cdot p_T$
	Medium	$30 \leq p_T < 170\,\text{GeV}$	$0.13 + 1.48 \cdot 10^{-3} \cdot p_T - 1.00 \cdot 10^{-5} \cdot p_T^2 + 2.65 \cdot 10^{-8} \cdot p_T^3 - 2.36 \cdot 10^{-11} \cdot p_T^4$
		$170 \leq p_T$	0.20
	Tight	$30 \leq p_T < 240\,\text{GeV}$	$0.024 + 5.27 \cdot 10^{-4} \cdot p_T - 3.72 \cdot 10^{-6} \cdot p_T^2 + 9.87 \cdot 10^{-9} \cdot p_T^3 - 8.83 \cdot 10^{-12} \cdot p_T^4$
		$240 \leq p_T$	0.044
light	Loose	$30 < p_T < 130\,\text{GeV}$	$0.124 - 1.0 \cdot 10^{-3} \cdot p_T + 1.06 \cdot 10^{-5} \cdot p_T^2 - 3.18 \cdot 10^{-8} \cdot p_T^3 + 3.13 \cdot 10^{-11} \cdot p_T^4$
		$130 \leq p_T$	$0.055 + 4.53 \cdot 10^{-4} \cdot p_T - 1.6 \cdot 10^{-7} \cdot p_T^2$
	Medium	$30 \leq p_T < 170\,\text{GeV}$	$9.59 \cdot 10^{-3} - 1.96 \cdot 10^{-5} \cdot p_T + 4.53 \cdot 10^{-7} \cdot p_T^2 - 1.08 \cdot 10^{-9} \cdot p_T^3 + 7.62 \cdot 10^{-13} \cdot p_T^4$
		$170 \leq p_T$	$5.07 \cdot 10^{-3} + 6.02 \cdot 10^{-5} \cdot p_T - 2.3 \cdot 10^{-8} \cdot p_T^2$
	Tight	$30 \leq p_T < 130\,\text{GeV}$	$1.24 \cdot 10^{-3} - 1.27 \cdot 10^{-5} \cdot p_T + 1.98 \cdot 10^{-7} \cdot p_T^2 - 7.46 \cdot 10^{-10} \cdot p_T^3 + 8.35 \cdot 10^{-13} \cdot p_T^4$
		$130 \leq p_T$	$1.08 \cdot 10^{-3} + 3.54 \cdot 10^{-6} \cdot p_T$

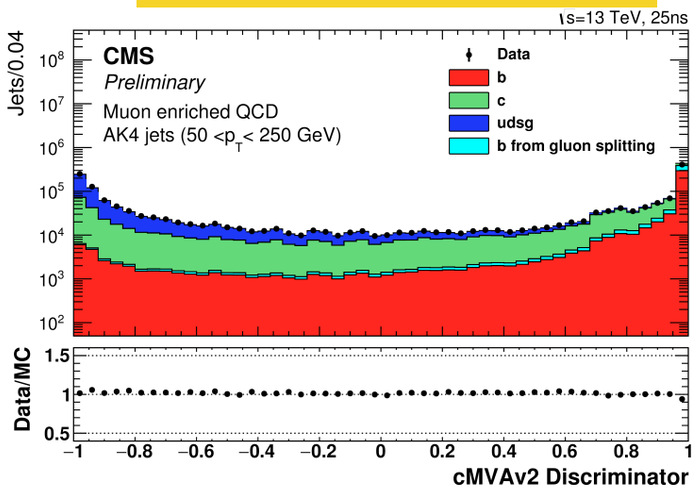
Parametrized b tag efficiencies are available in the PAS

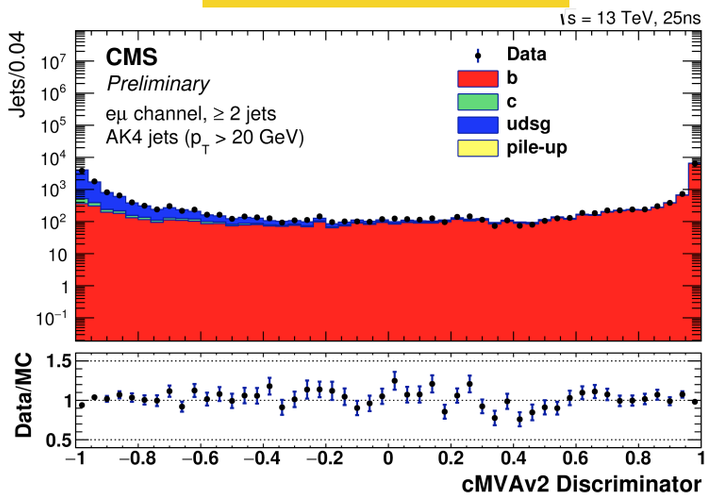
Commissioning for AK4 b jets

QCD inclusive | **QCD muon-enriched** | **ttbar inclusive**

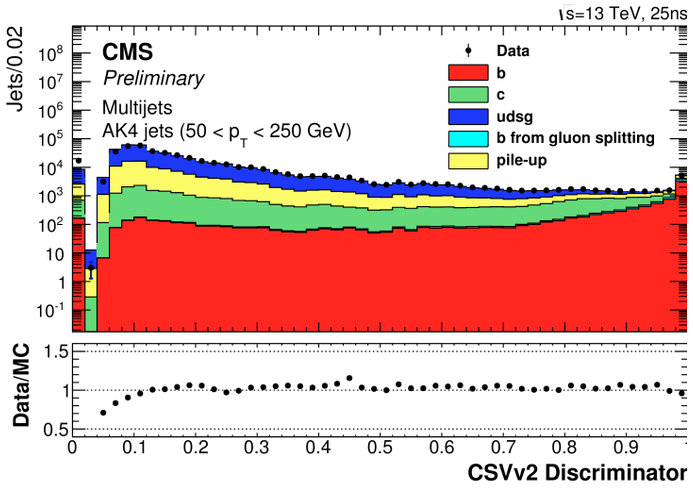
cMVAv2

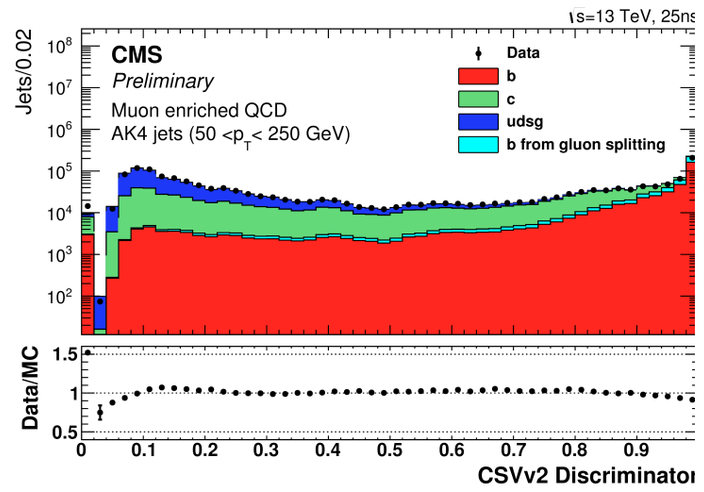


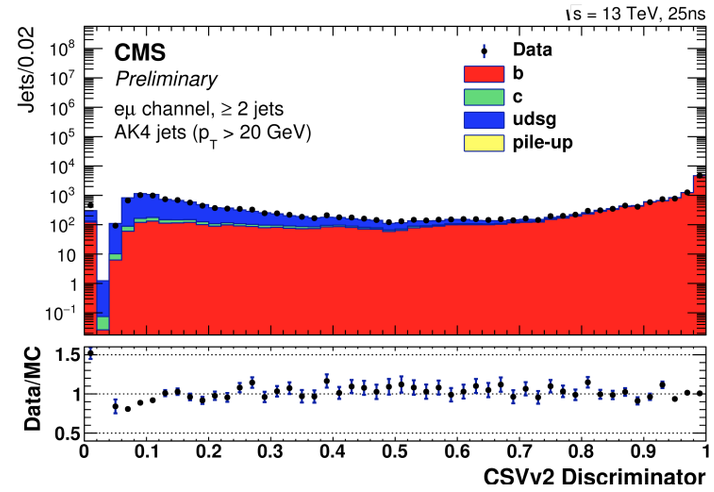




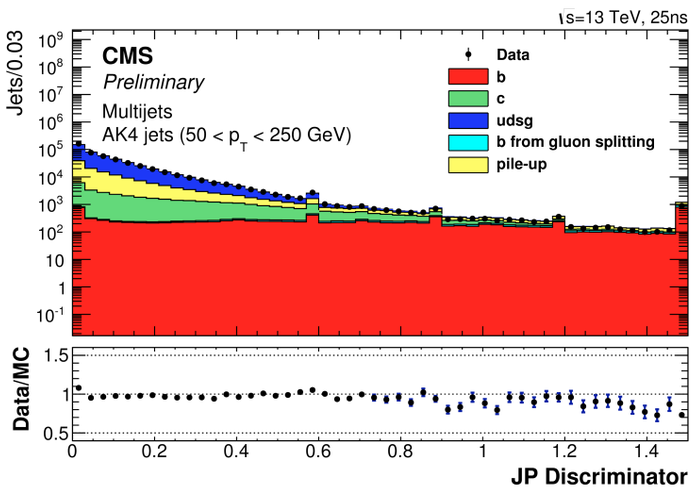
CSVv2

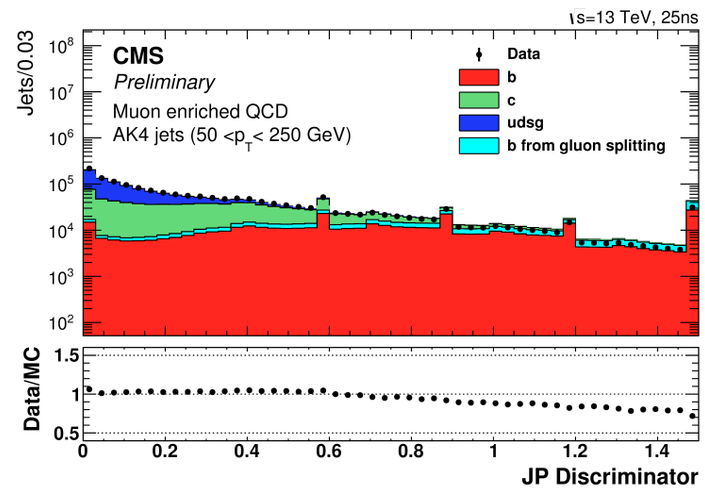


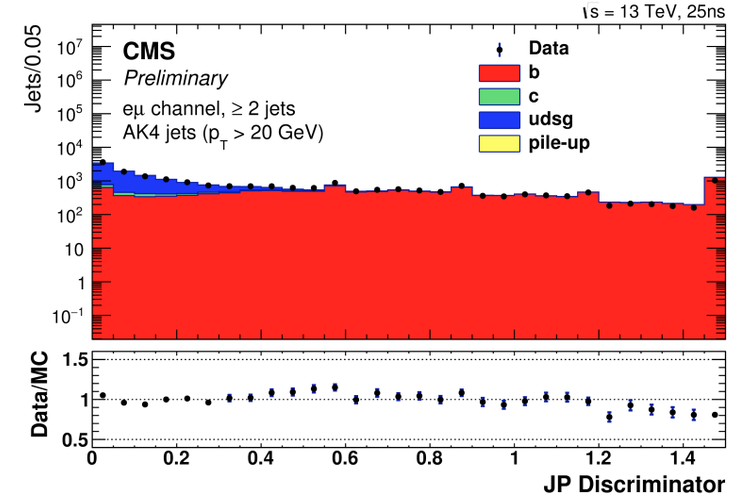




JP



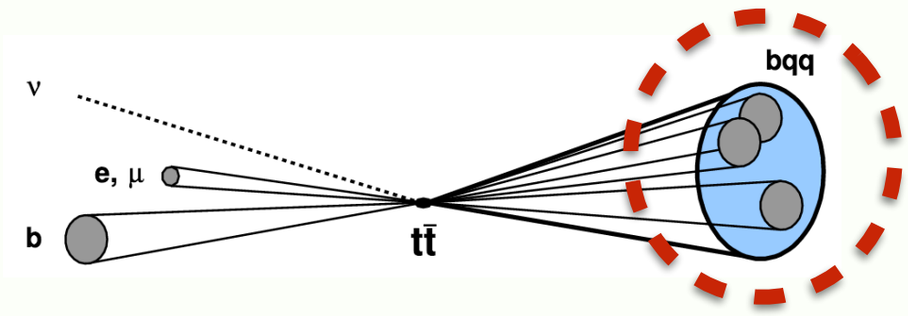




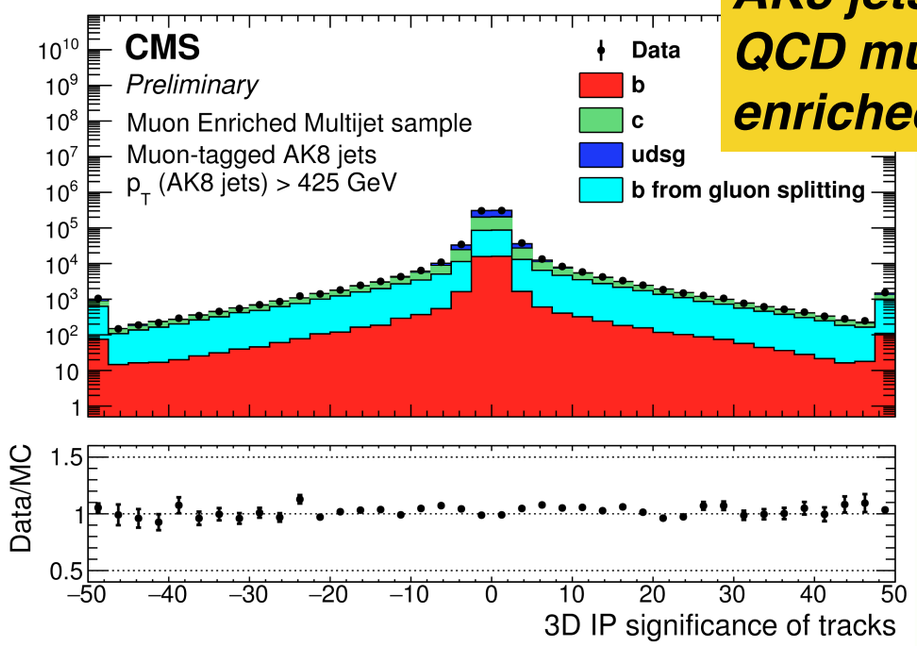
Commissioning in boosted topologies

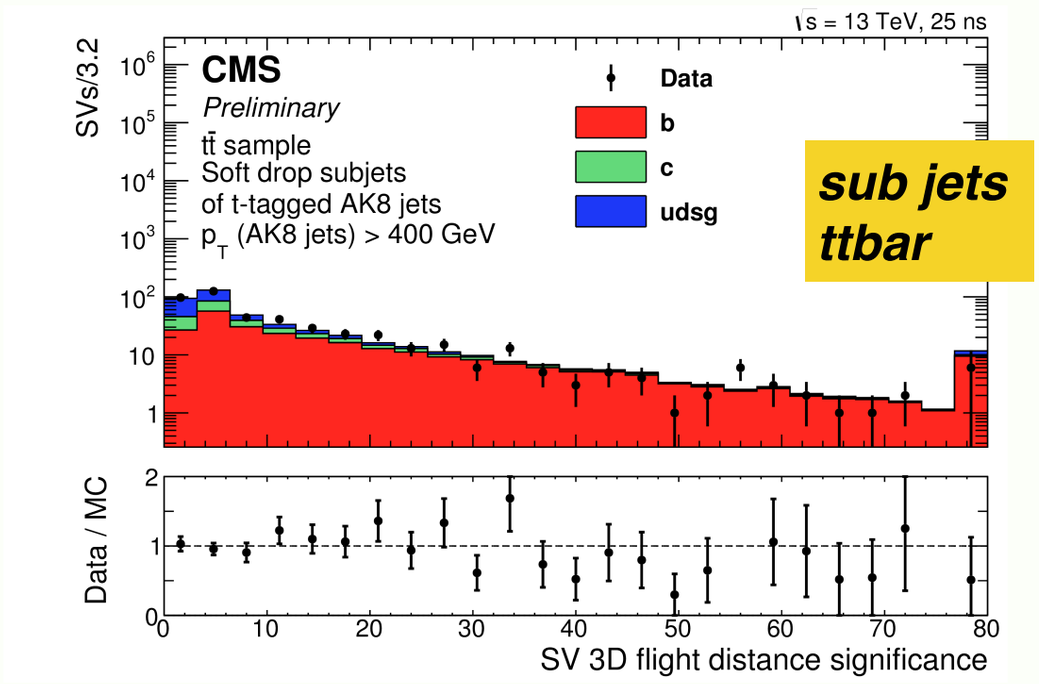
Check data/MC agreement for AK8 and AK4 soft drop sub jets

QCD muon-enriched (gluon splitting)
and **ttbar** (boosted hadronic top quarks)

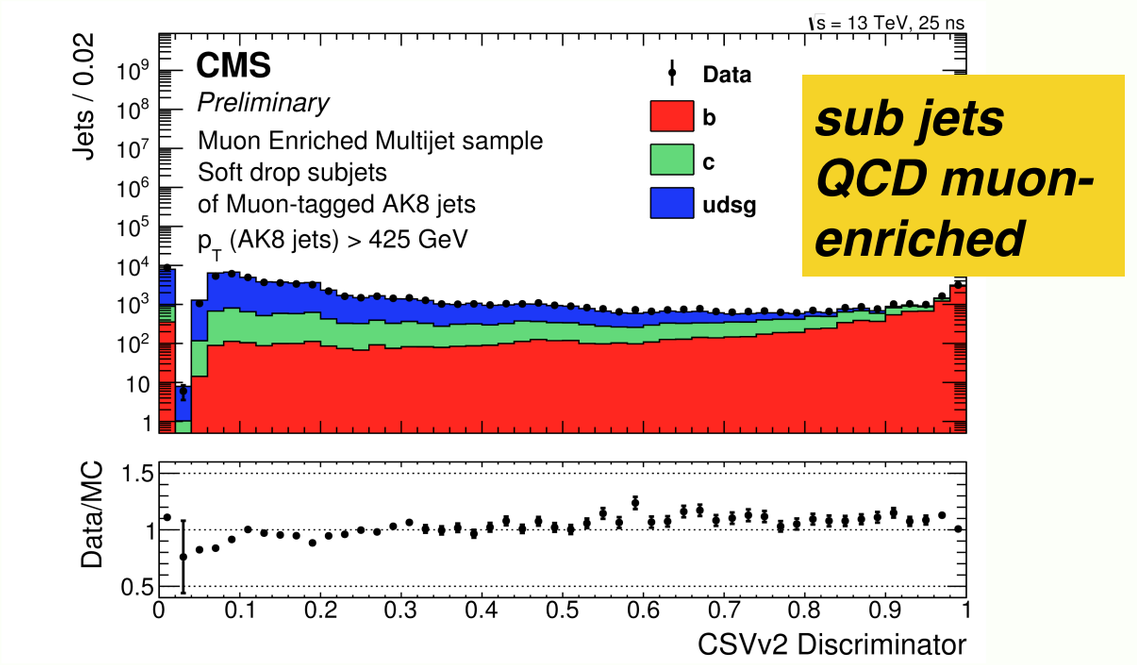


AK8 jets QCD muon-enriched





sub jets ttbar



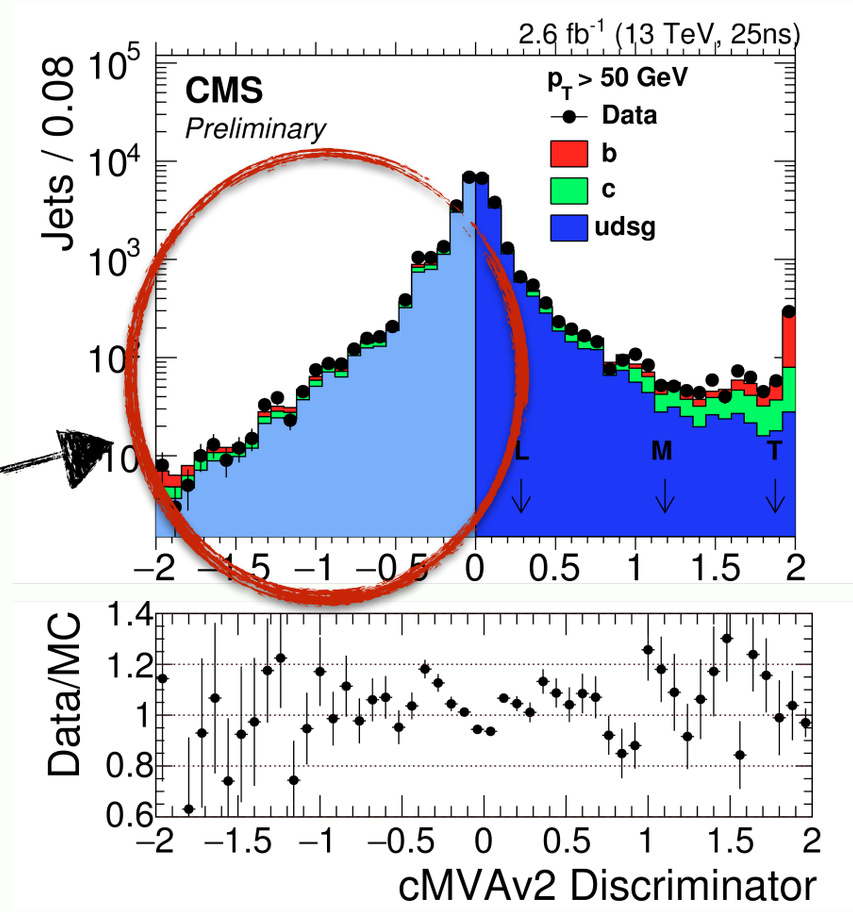
sub jets QCD muon-enriched

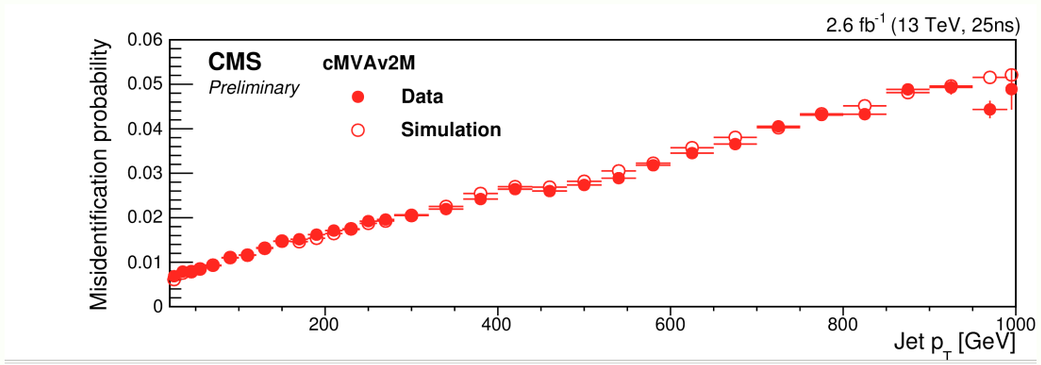
Mistag rate measurement

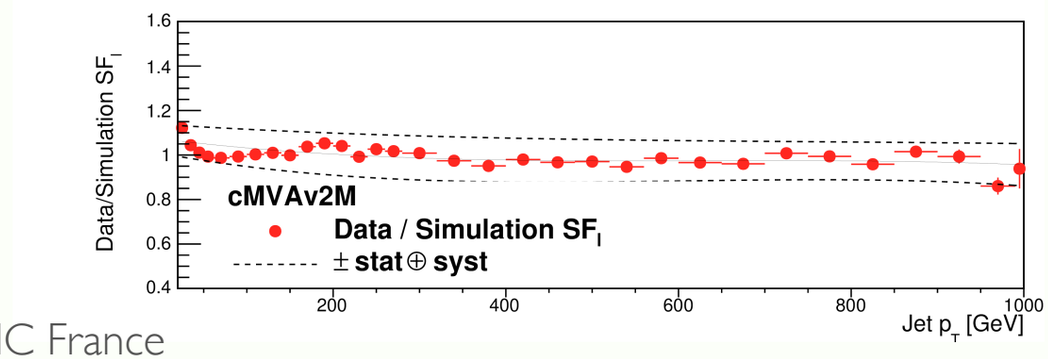
- **Mistag rate** ($\epsilon_{\mathrm{inc}}^{\mathrm{neg}}$) - efficiency to tag udsg as b jets
- Due to finite resolution of the inner detector, displaced vertices from long-lived particles and material interactions
- A ***negative tag method*** - use jets with negative impact parameter tracks
- Correct for b/c jet contamination and long-lived particles

$$k_{\mathrm{ll}} = \epsilon_l / \epsilon_l^{\mathrm{neg}} \qquad k_{\mathrm{hf}} = \epsilon_l^{\mathrm{neg}} / \epsilon_{\mathrm{inc}}^{\mathrm{neg}}$$

$$\epsilon_l = \epsilon_{\mathrm{inc}}^{\mathrm{neg}} k_{\mathrm{hf}} k_{\mathrm{ll}}$$

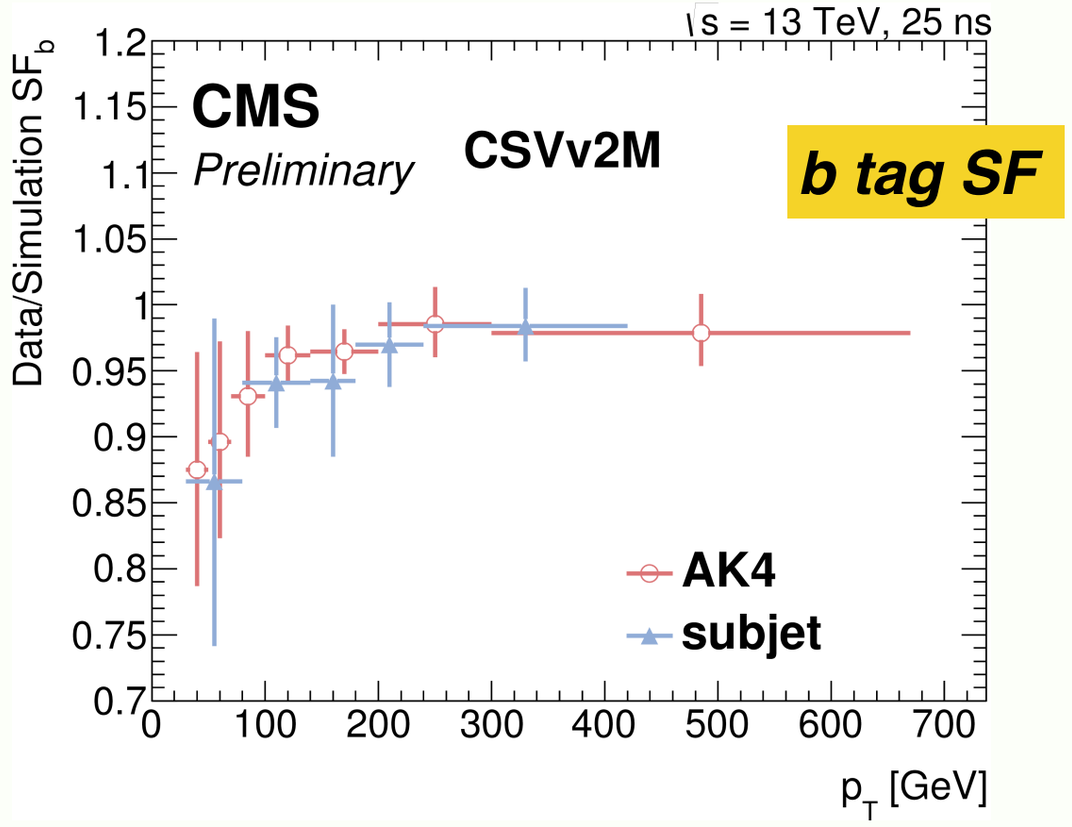


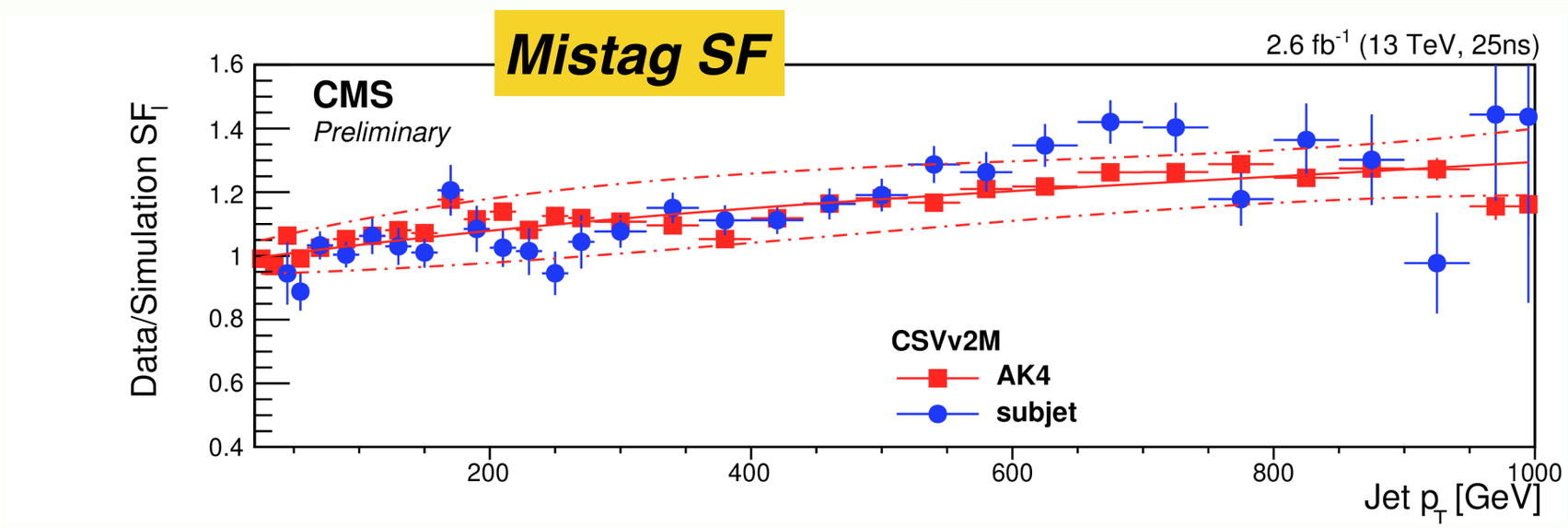




b tag calibration in boosted topologies

- **b tag efficiency** → QCD muon-enriched
- **Mistag rate** → QCD inclusive
- Measurement is performed on AK4 sub jets reconstructed within AK8 fat jets



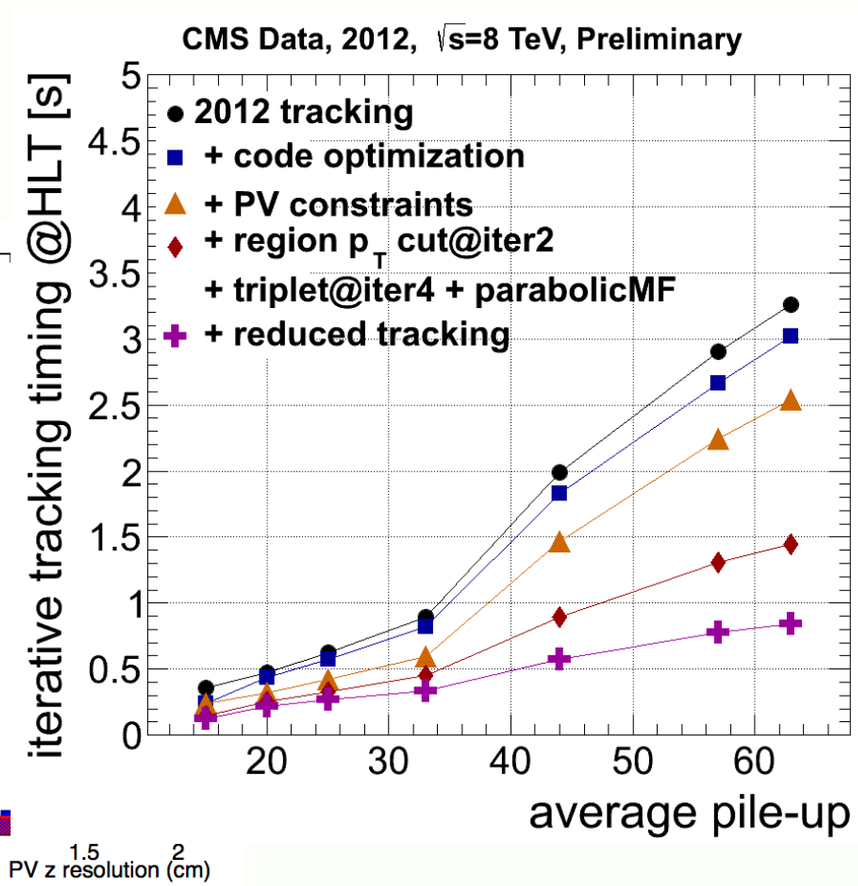


b tagging sequence at HLT in CMS

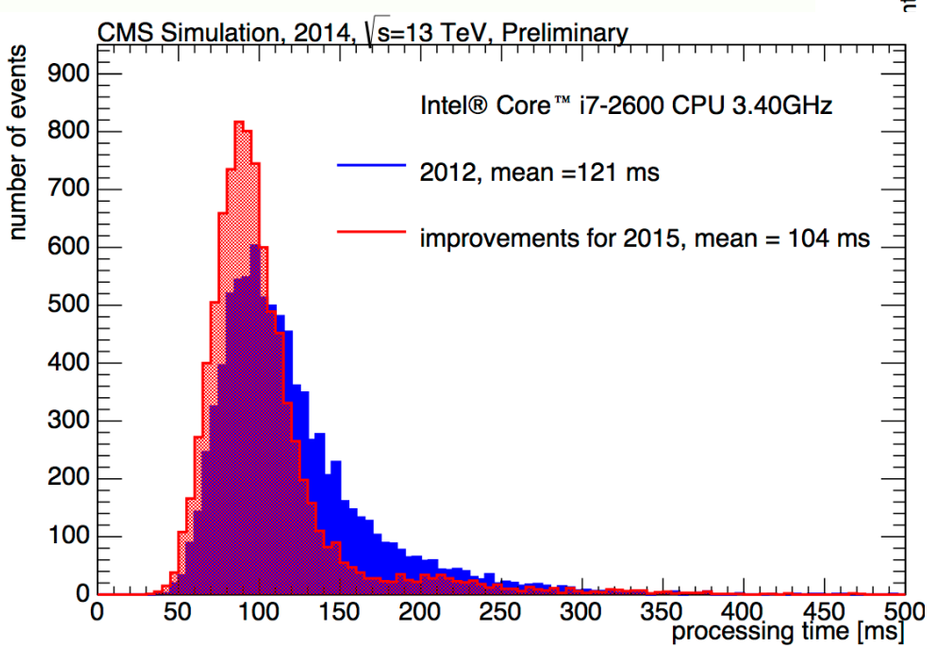
- ***Regional*** reconstruction of **pixel clusters** compatible with L1 calo jets
- Fast Primary Vertex (FastPV) reconstruction using only pixel information [$\sigma_z \approx 2$ mm]
- ***Regional*** **pixel track** reconstruction
- **Primary vertex** from pixel tracks [$\sigma \approx 100$ μm]
- Pixel+Strips **full track** reconstruction using ***iterative tracking*** [$\sigma \approx$ 20-30 μm]
- Apply CSVv2 ***b tagging*** algorithm

Also: L1 stage-2 trigger added in 2016 (topological selection at L1)

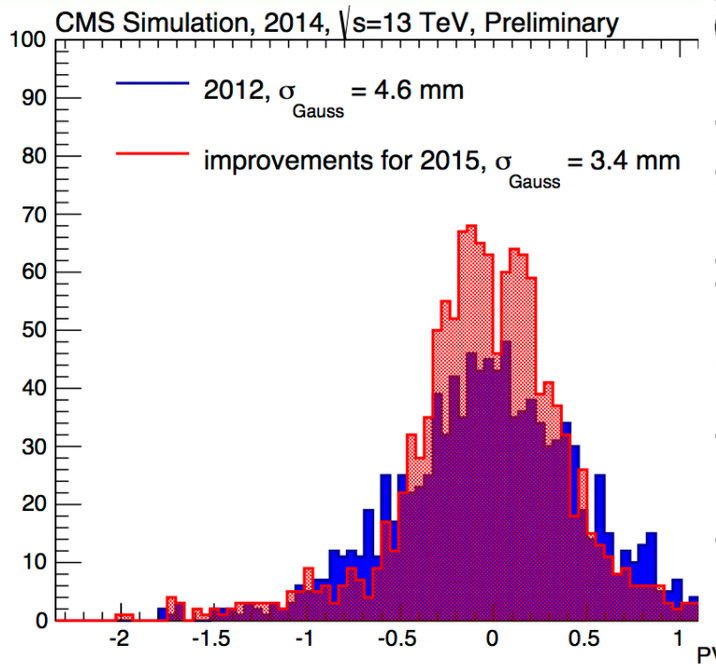
Iterative tracking performance with respect to nPV



Timing performance for Z(vv)H(bb) trigger path



FastPV resolution



b tag calibration in QCD events

- **System 8** method is based on extracting b-tagging efficiency from a ***system of 8 non-linear equations***
- Equations constructed from different b-tag samples (n, p, pTrel) defined by the reference and complementary b-tag selections
- Numerical methods are used to find a solution

$$n = n_b + n_{cl}$$
$$p = p_b + p_{cl}$$

b-tagging efficiency

$$n^{tag} = \varepsilon_b^{tag} n_b + \varepsilon_{cl}^{tag} n_{cl}$$
$$p^{tag} = \beta_{12}\varepsilon_b^{tag} p_b + \alpha_{12}\varepsilon_{cl}^{tag} p_{cl}$$

$$n^{p_{Trel}} = \varepsilon_b^{p_{Trel}} n_b + \varepsilon_{cl}^{p_{Trel}} n_{cl}$$
$$p^{p_{Trel}} = \beta_{23}\varepsilon_b^{p_{Trel}} p_b + \alpha_{23}\varepsilon_{cl}^{p_{Trel}} p_{cl}$$

$$n^{tag,p_{Trel}} = \beta_{13}\varepsilon_b^{tag}\varepsilon_b^{p_{Trel}} n_b + \alpha_{13}\varepsilon_{cl}^{tag}\varepsilon_{cl}^{p_{Trel}} n_{cl}$$
$$p^{tag,p_{Trel}} = \beta_{123}\varepsilon_b^{tag}\varepsilon_b^{p_{Trel}} p_b + \alpha_{123}\varepsilon_{cl}^{tag}\varepsilon_{cl}^{p_{Trel}} p_{cl}$$

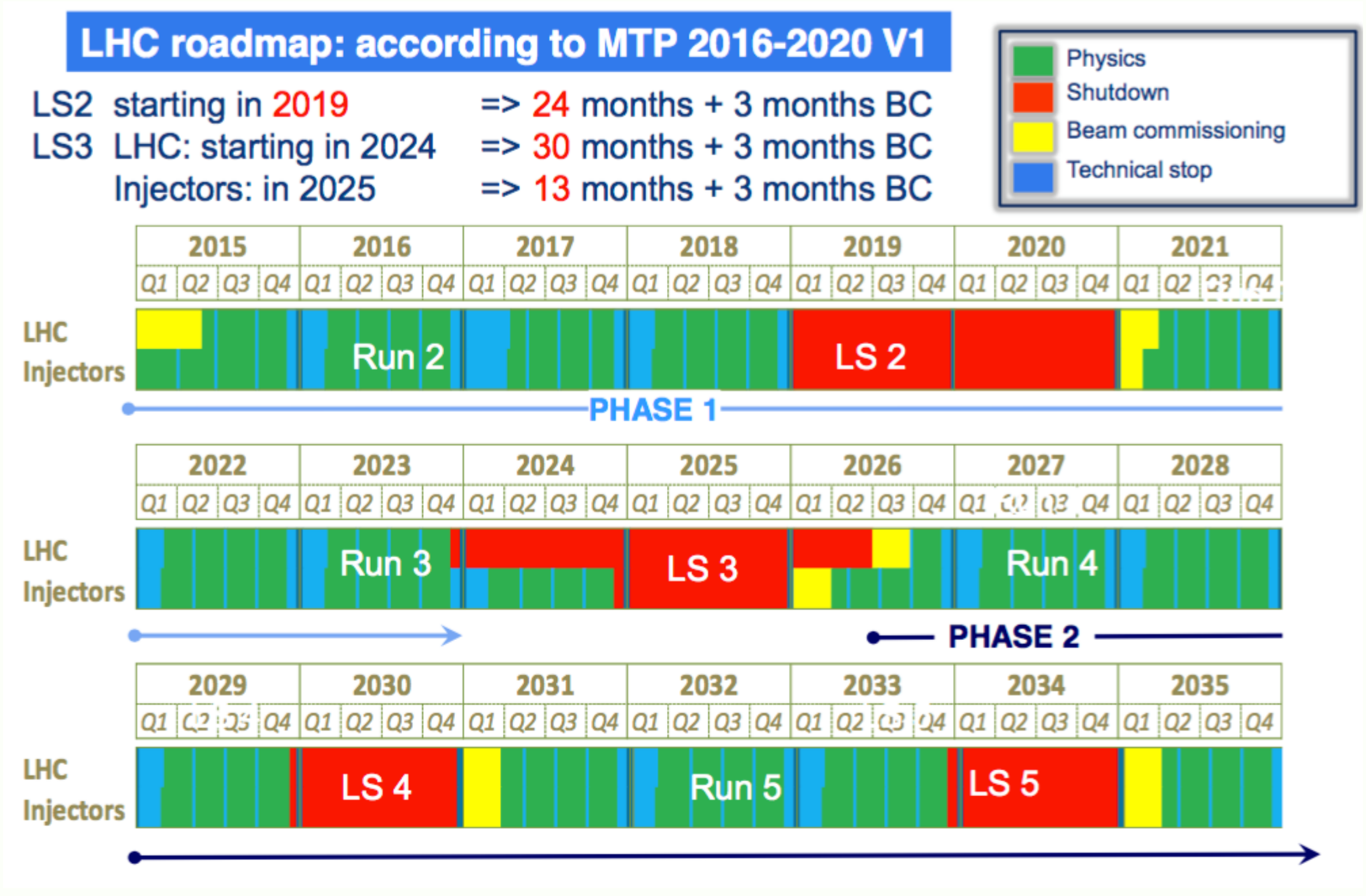
Unknowns

$$n_b, n_{cl}, p_b, p_{cl}$$
$$\varepsilon_b^{tag}, \varepsilon_{cl}^{tag}, \varepsilon_b^{p_{Trel}}, \varepsilon_{cl}^{p_{Trel}}$$

$$\alpha_{12}, \alpha_{23}, \alpha_{13}, \alpha_{123}$$
$$\beta_{12}, \beta_{23}, \beta_{13}, \beta_{123}$$

Correlation parameters

LHC roadmap



https://lhc-commissioning.web.cern.ch

Systematics in b tag efficiency measurement

- Several uncertainties affect the measurement of b tagging efficiency
  - ☑ Gluon splitting
  - ☑ b/c-quark fragmentation
  - ☑ Muon p_T
  - ☑ Away-jet tagger
  - ☑ c/l ratio
  - ☑ Selection on p_TRel
  - ☑ Difference between inclusive and muon jets
  - ☑ Generator uncertainties
    (PDF, parton shower, ISR/FSR, underlying event, B decay, etc.)
  - ☑ Pileup

source	size at ATLAS	size at CMS
b/c prod.	low pT: 0.1% - 0.2%, high pT for b-prod.: 1.2% - 2.0%	low pT: 0.1% - 0.3%, high pT: 0.5% - 1.3%
mu pT	first pT bin: 2.5%, 0.2% - 0.9% elsewhere	low pT: 0.1% - 1.1%, high pT: 0.1 - 0.9%
c/l ratio	<0.1% - 0.2%	<0.1% - 0.2%
b-frag	0.2% - 2.7%	0.2% - 0.8%
PS	0.1% - 1.5%	0.3% - 0.6%
IFSR	0.3% - 1.4%	0.3% - 0.6%

Treatment of correlations in b tagging uncertainties between ATLAS and CMS:
https://twiki.cern.ch/twiki/bin/view/LHCPhysics/BTaggingSystematics